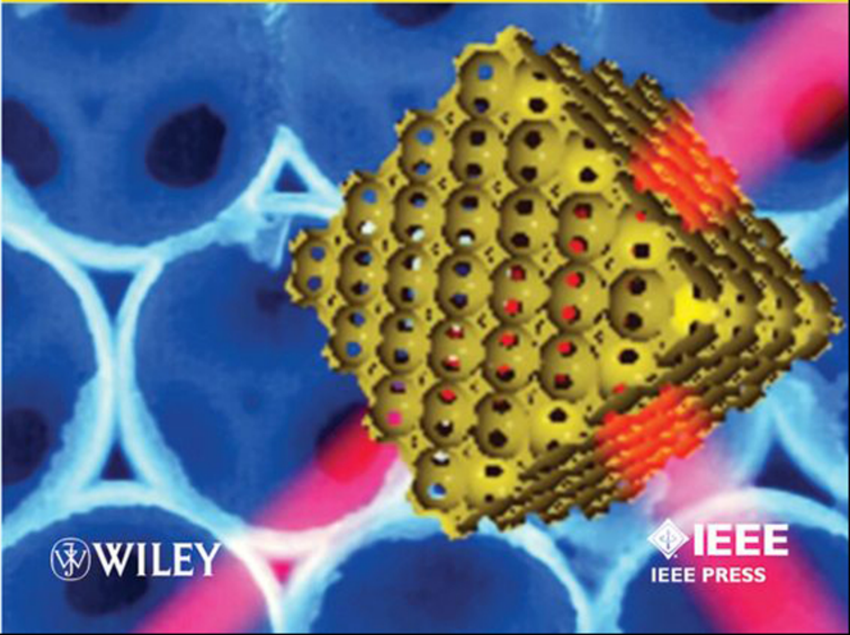
Future Trends in Microelectronics

From Nanophotonics to Sensors and Energy

Serge Luryi • Jimmy Xu • Alex Zaslavsky



WILEY
IEEE
IEEE PRESS

This page intentionally left blank

FUTURE TRENDS
IN MICROELECTRONICS

This page intentionally left blank

FUTURE TRENDS IN MICROELECTRONICS

From Nanophotonics to Sensors and Energy

Edited by

SERGE LURYI
JIMMY XU
ALEX ZASLAVSKY

IEEE

IEEE Press

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2010 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Future trends in microelectronics : from nanophotonics to sensors and energy / edited by Serge Luryi, Jimmy Xu, Alex Zaslavsky.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-55137-0 (cloth)
1. Microelectronics — Technological innovations. 2. Nanotechnology — Technological innovations. 3. Detectors —Technological innovations. I. Luryi, Serge. II. Xu, Jimmy. III. Zaslavsky, Alexander.
TK7874.F885 2010
621.381—dc22 2010006496

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface
S. Luryi, J. M. Xu, and A. Zaslavsky ix

1 OPTOELECTRONICS AND NANOPHOTONICS 1

Nanophotonics for Information Systems 5
Y. Fainman

What Will Modern Photonics Contribute to the Development of Future Optical Communication Technology? 23
Djafar K. Mynbaev

Ultrafast Nanophotonic Devices For Optical Interconnects 43
N. N. Ledentsov

Intersubband Quantum-Box Lasers: Progress and Potential as Uncooled Mid-Infrared Sources 49
D. Botez, G. Tsvid, M. D'Souza, J. C. Shin, Z. Liu, J. H. Park, J. Kirch, L. J. Mawst, M. Rathi, T. F. Kuech, I. Vurgaftman, J. Meyer, J. Plant, G. Turner, and P. Zory

GaSb-based Type-I Laser Diodes Operating at 3 μm and Beyond 65
L. Shterengas, G. Kipshidze, T. Hosoda, and G. Belenky

Bridging Optics and Electronics with Quantum Cascade Lasers, Antennas, and Circuits 75
J. Faist, C. Walther, M. Amanti, G. Scalari, M. Fischer, and M. Beck

Towards Intersubband Polaritonics: How Fast Can Light and Electrons Mate? 85
A. A. Anappara, L. Sorba, A. Tredicucci, G. Günter, A. Sell, A. Leitenstorfer, R. Huber, S. De Liberato, C. Ciuti, and G. Biasiol

Si_3N_4/SiO_2 Planar Photonic Structures Fabricated by Focused Ion Beam 97
L. A. M. Barea, F. Vallini, D. L. S. Figueira, A. Da Silva Filho, N. C. Frateschi, and A. R. Vaz

2 ELECTRONIC DEVICES AND SYSTEMS 105

Silicon-Based Devices and Materials for Nanoscale CMOS and Beyond-CMOS 109
F. Balestra

Device Proposals Beyond Silicon CMOS 127
P. M. Solomon

GeOI as a Platform for Ultimate Devices 141
W. Van Den Daele, S. Cristoloveanu, E. Augendre, C. Le Royer, J.-F. Damlencourt, D. Kazazis, and A. Zaslavsky

Simulation of Self-Heating Effects in Different SOI MOS Architectures 159
Enrico Sangiorgi, Marco Braccioli and Claudio Fiegna

Nanowires: Technology, Physics and Perspectives 171
D. Grützmacher, Th. Schäpers, S. Mantl, S. Feste, Q. T. Zhao, H. Hardtdegen, R. Calarco, M. Lepsa, and N. Demarina

Emerging Nanotechnology for Integration of Nanostructures in Nanoelectronic Devices 183
T. Baron, C. Agraffeil, F. Dhalluin, M. Kogelschtaz, G. Cunge, T. Chevolleau, B. Salem, B. Salhi, H. Abed, A. Potié, L. Latu-Romain, C. Ternon, K. Aissou, L. Montès, Mur, G. Molas, B. De Salvo, E. Jalaguier, T. Ernst, P. Ferret, P. Gentile, and N. Pauc

Scrolled Si/SiGe Heterostructures as Building Blocks for Tube-Like Field-Effect Transistors 191
N. V. Demarina and D. Grützmacher

Silicon Nanowire-Based Nonvolatile Memory Cells: Progress and Prospects 201
Qiliang Li, X. Zhu, Y. Yang, D. E. Ioannou, J. S. Suehle, and C. A. Richter

Prospects and Challenges of Next-Generation Flash Memory Devices 211
Jang-Sik Lee

Chalcogenide Glassy Semiconductors – Could They Replace Silicon in Memory Devices? 219
K. D. Tsendin

Current Status and Recent Developments in RSFQ Processor Design 229
M. Dorojevets

1/*f* Noise: The Funeral is Cancelled (or Postponed) 239
M. E. Levinshtein and S. L. Rumyantsev

3 PHYSICS, BIOLOGY, AND OTHER SISTER SCIENCES 247

Spin Hall Effect 251
M. I. Dyakonov

Can Biology Provide Creative Solutions for Next-Generation Memory Devices? 265
W. E. van den Berg and S. A. Kushner

Spin Screening of Magnetization Due to Inverse Proximity Effect in Superconducting/Ferromagnetic Bilayers 273
V. Shelukhin, M. Karpovski, A. Palevski, J. Xia,
A. Kapitulnik, and A. Tsukernik

Silicon for Spintronic Applications: Strain-Enhanced Valley Splitting 281
V. Sverdlov, O. Baumgartner, T. Windbacher, and S. Selberherr

Graphene-Based Terahertz Devices: Concepts and Characteristics 293
V. Ryzhii, M. Ryzhii, A. Satou, N. Ryabova, T. Otsuji, V. Mitin,
F. T. Vasko, A. A. Dubinov, V. Y. Aleshkin, and M. S. Shur

Directed Self-Assembly – A Controllable Route to Optical and Electronic Devices Based on Single Nanostructures 307
R. L. Williams, D. Dalacu, M. E. Reimer, K. Mnaymneh, V. Sazonova,
P. J. Poole, G. C. Aers, R. Cheriton, S. Frédérick, D. Kim,
J. Lapointe, P. Hawrylak, and M. Korkusiński

4 SENSORS, DETECTORS, AND ENERGY 317

Three-Dimensional Position-Sensitive Wide Bandgap Semiconductor Gamma-Ray Imaging Detectors 321
Zhong He

Semiconductor Scintillator for Three-Dimensional Array of Radiation Detectors 331
Serge Luryi and Arsen Subashiev

Semiconductor Gamma Radiation Detectors: Band Structure Effects in Energy Resolution 347
Arsen Subashiev and Serge Luryi

The Future of Microelectronics is ... Macroelectronics 365
M. A. Alam, N. Pimparkar, and B. Ray

An Integration Challenge: Information and Communication Technologies to Address Indoor Air Quality in Commercial Buildings 377
M. D'Iorio

Quantum-Dot Infrared Photodetectors: In Search of the Right Design for Room-Temperature Operation 385
A. Sergeev, V. Mitin, L. H. Chien, and N. Vagidov

Treating the Case of Incurable Hysteresis in VO_2 395
M. Gurvitch, S. Luryi, A. Polyakov, and A. Shabalov

Exploratory Studies on Silicon-Based Oxide Fuel Cell Power Sources Incorporating Ultrathin Nanostructured Platinum and Cerium Oxide Films as Anode Components 411
Bo-Kuai Lai, A. C. Johnson, H. Xiong, C. Ko, and S. Ramanathan

INDEX 423

Preface

S. Luryi
Dept. of Electrical and Computer Engineering
SUNY-Stony Brook, Stony Brook, NY 11794, U.S.A.

J. M. Xu and A. Zaslavsky
Division of Engineering, Brown University, Providence, RI 02912, U.S.A.

This book is a brainchild of the sixth workshop in the *Future Trends in Microelectronics* series (FTM-6). The first of the FTM conferences, "*Reflections on the Road to Nanotechnology*", had gathered in 1995 on Ile de Bendor, a beautiful little French Mediterranean island.[1] The second FTM, "*Off the Beaten Path*", took place in 1998 on a larger island in the same area, Ile des Embiez.[2] Instead of going to a still larger island, the third FTM, "*The Nano Millennium*" went back to its origins on Ile de Bendor in 2001.[3] To compensate, the next FTM, "*The Nano, the Giga, the Ultra, and the Bio*" took place on the biggest French Mediterranean island of them all, Corsica.[4] Normally, the FTM workshops gather every three years; however, the FTM-4 was held one year ahead of the usual schedule, in the summer of 2003, as a one-time exception. Continuing its inexorable motion eastward, the fifth FTM workshop, "*Up the Nano Creek*", had convened on Crete, Greece, in June of 2006.[5] The inexorable motion was then interrupted to produce a semblance of a random walk in the Mediterranean and the last workshop, FTM-6 "*Unmapped Roads*" went to the great Italian island of Sardinia (June 2009).

The FTM workshops are relatively small gatherings (less than 100 people) by invitation only. If you, the reader, wish to be invited, please consider following a few simple steps outlined on the conference website. The FTM website at *www.ece.sunysb.edu/~serge/FTM.html* contains links to all past and planned workshops in the series, their programs, publications, sponsors, and participants. Our attendees have been an illustrious lot. Suffice it to say that among FTM participants we find five Nobel laureates (Zhores Alferov, Herbert Kroemer, Horst Stormer, Klaus von Klitzing, and Harold Kroto) and countless others poised for a similar distinction. To be sure, high distinction is not a prerequisite for being invited to FTM, but the ability and desire to bring fresh ideas is. All participants of FTM-6 can be considered authors of this book, which in this sense is a collective treatise.

The main purpose of FTM workshops is to provide a forum for a free-spirited exchange of views, projections, and critiques of current trends and future directions, among the leading professionals in industry, academia, and government. It is a common view among the leading professionals in microelectronics, that its current explosive development will likely lead to profound paradigm shifts in the near future. Identifying the plausible scenarios for the future evolution of microelectronics presents a tremendous opportunity for constructive action today.

For better or worse our civilization is destined to be based on electronics. Ever since the invention of the transistor and especially after the advent of integrated circuits, semiconductor devices have kept expanding their role in our lives. Electronic circuits entertain us and keep track of our money, they fight our wars and decipher the secret codes of life, and one day, perhaps, they will relieve us from the burden of thinking and making responsible decisions. Inasmuch as that day has not yet arrived, we have to fend for ourselves. The key to success is to have a clear vision of where we are heading.

Some degree of stability is of importance in these turbulent times and should be welcome. Thus, although the very term "*microelectronics*" has been generally re-christened "*nanoelectronics*", we have stuck to the original title of the FTM workshop series.

The present volume contains a number of original papers, some of which were presented at FTM-6 in oral sessions, other as posters. From the point of view of the program committee, there is no difference between these types of contributions in weight or importance. There was, however, a difference in style and focus – and that was intentionally imposed by the organizers. All speakers were asked to focus on the presenter's views and projections of future directions, assessments or critiques of important new ideas/approaches, and *not* on their own achievements. This latter point is perhaps the most distinctive feature of FTM workshops. Indeed, we are asking scientists not to speak of their own work! This has proven to be successful, however, in eliciting powerful and frank exchange. The presenters are asked to be provocative and/or inspiring. Latest advances made and results obtained by the participants are to be presented in the form of posters and group discussions.

Each day of the workshop was concluded by an evening panel or poster session that attempted to further the debates on selected controversial issues connected to the theme of the day. Each such session was chaired by a moderator who invited two or three attendees of his or her choice to lead with a position statement, with all other attendees serving as panelists. The debate was forcefully moderated and irrelevant digressions cut off without mercy. Moderators were also assigned the hopeless task of forging a consensus on critical issues.

All FTM workshops adhered to these principles in the past and, hopefully, will do so in the future. To accommodate these principles, the FTM takes a format that is less rigid than usual workshops to allow and encourage uninhibited exchanges and sometimes confrontations of different views. A central theme is designed together with the speakers for each day. Another traditional feature of FTM workshops is a highly informal vote by the participants on the relative importance of various fashionable current topics in modern electronics research. This tradition owes its origin to Horst Stormer, who composed the original set of questions and maintained the results over four conferences. These votes are perhaps too bold and irreverent for general publication, but they are carefully maintained and made available to every new generation of FTM participants. Unfortunately, Horst missed the Sardinia gathering, but the tradition was maintained in his absence. Another traditional vote concerned the best poster. The 2009 winning poster was "Heterogeneous integration of nanowires and nanotubes on CMOS" by Sameer Sonkusale.

From all the deliberations and discussion at FTM-6 the following trends could be discerned, with the caveat that our crystal ball is as muddy as ever.

Firstly, although silicon is undoubtedly still full of steam, the word "post-CMOS" has become a commonplace. Several FTM-6 presentations included the expression "beyond silicon" in the title. A clearly discernible trend is the quest for novel and exotic materials, as well as utilization of nanoscale phenomena. It looks like we are back to fundamentals. The big-brother silicon is clearly pressed (at least in terms of the conference publicity) by its much nimbler sibling carbon, which is capable of producing graphene sheets with miraculous properties. Germanium is another column-IV element that is challenging silicon at its own game. Away from column-IV materials, we can single out an exciting report describing the *superconducting* processors and systems based on the rapid single flux quantum (RSFQ) circuit technology.

Among other – perhaps niche – materials, discussed at the workshop, we find chalcogenide glassy semiconductors for memory applications and vanadium dioxide that undergoes semiconductor-to-metal transition above room temperature. It turns out that in VO_2 films the hysteresis of this transition can be tamed to yield reproducible linear excursions suitable for bolometric applications.

Secondly, a sizeable part of the discussion at FTM-6 was devoted to optoelectronics, most notably to nanophotonics issues, such as photonic crystals, plasmonics, and terahertz cascade lasers. Several authors discussed applications of nanophotonics in optical interconnects and communications. This topic is traditional for FTM but the emphasis on nanophotonics is a discernible new trend.

Finally, a noteworthy fraction of the discussion at our microelectronics workshop dealt with *macro* electronics, such as flat panel displays, solar cells, and bioelectronics sensors. For such systems, the guiding principles are cost (the smaller the better) and often the size (the larger the better!). A new topic for FTM was the discussion of high-energy radiation sensors. Development of gamma detectors, capable of isotope discrimination and the determination of the direction to source, is a very timely challenge, especially in homeland security applications. The FTM-6 workshop had included a mini-symposium on gamma radiation sensors, and three papers presented at that symposium have made it into this volume. The purpose of the gamma-radiation symposium was to acquaint the traditional FTM audience with the basic issues of radiation detection, in hope of recruiting new blood to address issues in the nuclear-engineering field that have become timely in the face of the new challenge. Presentations at the symposium were therefore somewhat pedagogical in nature.

Not every contribution presented at FTM-6 has made it into this book (not for the lack of persistence by the editors). We sorely miss the exciting contribution by

Mark Pinto, a "captain of industry" from Applied Materials, who spoke on large-scale photovoltaics. Abstracts of his and all other presentations can be found on the workshop program webpage, *http://www.ee.sunysb.edu/~serge/ARW-6/ program.html*

The FTM meetings are known for the professional critiques – or even demolitions – of fashionable trends, that some may characterize as hype. The previous workshops had witnessed powerful assaults on quantum computing, molecular electronics, and spintronics. The majority of FTM participants did not consider quantum computing a realistic future technology, but gave it credit as an interesting playground for physicists with some hope of settling old debates about the wavefunction collapse and other fundamental issues. It seems that by now most of the hype associated with quantum computing has dissipated and perhaps we can take some credit for the more balanced outlook that has emerged since.

Another characteristic of FTM meetings is the settling of scientific bets, a tradition that dates back to the FTM-2 wager between Nikolai Ledentsov (for) and Horst Stormer (against) the coming dominance of quantum dot-based lasers – a bet that Horst collected in 2004, at FTM-4. At FTM-6, the two bets coming due were on the relative strength of personal computer-based chess programs *versus* the human world champion (the bettors were Serge Luryi for the humans and Alex Zaslavsky for the PCs) and the predicted capture of half of the Si technology market by SiGe devices (the bettors were Erich Kasper for SiGe and Sorin Cristoloveanu against). Confronted with the evidence of a modern PC beating grandmasters even while starting a pawn down, Serge conceded at the workshop, whereas Erich had to concede long-distance. Yet another bet on the putative future dominance of SOI technology was made at FTM-6, to be adjudicated at a future workshop.

We have grouped all contributions into four chapters, entitled *I. Opto-electronics and nanophotonics*; *II. Electronic devices and systems*; *III. Physics, biology, and other sister sciences*; and *IV. Sensors, detectors, and energy*. The breakdown could not be uniquely defined, because some papers fit two or even three categories! To produce a coherent collective treatise out of all of this, the interaction between FTM participants had begun well before their gathering at the workshop. All the proposed presentations were posted on the web in advance and could be subject to change up to the last minute to take into account peer criticism and suggestions. After the workshop is over, these materials (not all of which have made it into this book) remain on the web indefinitely, and the reader can peruse them starting at the *www.ece.sunysb.edu/~serge/FTM.html* home page.

Acknowledgments

The 2009 FTM workshop on Sardinia and therefore this book were possible owing to support from:

- U.S. Department of Defense: ARO, DTRA, ONR;

- U.S. DoD European offices: EOARD, ONR–Global;
- NRC of Canada (Institute for Microstructural Sciences);
- Industry: Applied Materials Inc.;
- Academia: SUNY–Stony Brook, The World-Class University Program at the Seoul National University.

On behalf of all Workshop attendees sincere gratitude is expressed to the above organizations for their generous support.

Finally, the organizers wish to thank all of the contributors to this volume and all the attendees for making the workshop a rousing success.

References

1. S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology*, NATO ASI Series E Vol. 323, Dordrecht: Kluwer Academic, 1996.
2. S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: The Road Ahead*, New York: Wiley Interscience, 1999.
3. S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: The Nano Millennium*, New York: Wiley Interscience/IEEE Press, 2002.
4. S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: The Nano, The Giga, and The Ultra*, New York: Wiley Interscience/IEEE Press, 2004.
5. S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Up the Nano Creek*, New York: Wiley Interscience/IEEE Press, 2007.

This page intentionally left blank

Part I

Optoelectronics and Nanophotonics

This page intentionally left blank

1 Optoelectronics and Nanophotonics

Optoelectronics – an artful expansion of the electron degree of freedom to the mating of electrons with photons continued to be a major theme of discussion at the FTM-6 gathering. With a focus initially on fiber-optics technologies at the first FTM in anticipation of the ensuing leaps in telecom systems, the discussions have reached further and wider. FTM workshops always covered the persistent efforts on solar cells and photonic crystals, long before they became fashionable, but now the discussions featured a notable emphasis on nanophotonics. At the meeting, the projections of fascinating trends ranging from the very large scale (wall-sized) photovoltaics to the very localized photon traps in 3D photonic crystal optical chips and the very efficient light emitting diodes were presented by Mark Pinto, Sajeev John, Claude Weisbuch, and others. In this chapter, nanophotonics in optical interconnects and communications, sub-wavelength resonators, quantum-box and quantum-well light sources are featured.

Contributors

1.1 Y. Fainman

1.2 Djafar K. Mynbaev

1.3 N. N. Ledentsov

1.4 D. Botez, G. Tsvid, M. D'Souza, J. C. Shin, Z. Liu, J. H. Park, J. Kirch, L. J. Mawst, M. Rathi, T. F. Kuech, I. Vurgaftman, J. Meyer, J. Plant, G. Turner, and P. Zory

1.5 L. Shterengas, G. Kipshidze, T. Hosoda, and G. Belenky

1.6 J. Faist, C. Walther, M. Amanti, G. Scalari, M. Fischer, and M. Beck

1.7 A. A. Anappara, L. Sorba, A. Tredicucci, G. Günter, A. Sell, A. Leitenstorfer, R. Huber, S. De Liberato, C. Ciuti, and G. Biasiol

1.8 L. A. M. Barea, F. Vallini, D. L. S. Figueira, A. Da Silva Filho, N. C. Frateschi, and A. R. Vaz

This page intentionally left blank

Nanophotonics for Information Systems

Y. Fainman
Dept. of Electrical and Computer Engineering, UC–San Diego
9500 Gilman Drive, La Jolla, California, 92093-0407, U.S.A.

1. Introduction

Optics has the potential to solve some of the most pressing problems in communication and computing hardware. It promises crosstalk-free interconnects with essentially unlimited bandwidth; long-distance data transmission without skew and without power- and time-consuming regeneration; miniaturization; parallelism; and efficient implementation of important algorithms such as Fourier transforms. In the past, when the speed of digital computers was able to support only relatively small information processing throughput, optical information processing techniques were developed and used to construct processors and systems in support of numerous applications that required high throughput for real time operation. These methods exploited the parallelism of optics supported by the richness of the modal continuum of free space and a variety of optoelectronic devices that were developed in support of these applications and systems. The constructed information processing systems and concepts were used for image processing,[1-4] pattern recognition,[5,6] neural networks,[7] and linear algebra calculus[8] – to name a few. However with rapid advancements of the speed and, therefore, the information processing throughput of digital computers, the optical signal processing systems were not able to support these applications in a broad sense due to high cost, lesser accuracy, and lack of user-friendly interfaces. Later, the optical information processing transformed from parallelism in space domain to parallelism in optical frequency domain in support of processing information carried by ultrashort optical pulses in the femtosecond range. Such waveforms vary too rapidly for even the fastest photodetectors to resolve, leading to the need to develop optical information processing methods. Time-domain and spectral-domain processors utilized linear[9] and nonlinear[10] processes, and found useful applications for ultrafast waveform synthesis, detection and processing.[10–12]

It is evident that optical processing in space and time has so far failed to move out of the lab. The free-space and guided-wave devices are costly, bulky, and fragile in their alignment. They are also difficult to integrate with electronic systems, both in terms of the fabrication process and in terms of delivery and retrieval of the massive volumes of data the optical elements can process. However, with the most recent emphases on construction of *chip-scale integration* using advanced lithographic tools employed in surrounding electronics, things may be changing. Experts predict lithographic resolution as fine as 16 nm by the year

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

2020, which is about a hundred times smaller than the telecommunication wavelength of 1550 nm. These techniques can be used to create deeply subwavelength features that act as *metamaterials* with optical properties controlled by the *density* and *geometry* of the pattern and its *constituent materials*. In this Section we focus on metamaterials composed primarily of dielectrics that are engineered on the nanometer scale so as to have emergent optical properties not otherwise present. The increased localization of the optical field as a result of these engineered materials brings about phenomena such as form-birefringence, structural dispersion and enhanced optical nonlinear interactions. Equivalently, characteristics such as the local polarizability and dispersion of the metamaterial may be controlled by geometry, properties of constituent materials and their composition. The introduction of periodicity into these engineered materials modifies the dispersion relations and can be used to create an artificial bandgap.[13–15] The manipulation and modification of this periodicity allows the bandgap to shift and parts of the bandgap to be accessed by propagating modes. Photonic crystal (PhC) waveguides rely on this exact concept – a line of defects is introduced into the otherwise periodic structure so as to guide light.[16–18] The confinement of light within the PhC lattice is also used to realize devices such as super collimators, super prisms, super lenses, omnidirectional filters, modulators and lasers through proper design and optimization of Bloch modes.[19–28]

Similar to the PhC in its periodic nature is a class of metamaterials that exploit the advantages of both continuous free-space and discrete guided-wave modes. The simplest example involves the propagation of light in a waveguiding slab, where confinement occurs only in the vertical direction; the free space propagation occurs in the plane of the slab. This configuration, aptly termed "free space optics on a chip" (FSOC), enables interaction with discrete optical components that are located along the propagation direction. Functionalizing devices for such integrated systems would require free space implementations of focusing, beam steering, and wavelength selectivity.[29] Realization of these functionalities can exploit periodic, quasi-periodic or even random nanostructured composites. By altering the surface morphology of a dielectric on the nanoscale via nanolithography and advanced etching techniques, we can realize these complex structures and control the materials' local polarizability. As we shall see in this chapter, these structures fall into the deep sub-wavelength regime with spatial features $\ll \lambda/2n$ and require metamaterial engineering with very high spatial resolution.

The engineering of composite dielectrics can continue to larger scales creating metamaterials that involve feature sizes on a larger sub-wavelength scale, *e.g.* just $< \lambda/2n$. The common themes of periodicity/quasi-periodicity and enhanced effects due to light confinement of guided modes will still remain. Continuing the simplification of PhC with periodicity in two dimensions that has been used for planar confinement, an alternative system whereby 2D light confinement is achieved by total internal reflection and 1D periodicity is introduced to create a bandgap in the third dimension. One method involves using a channel waveguide to guide light, rather than a slab waveguide as in the FSOC case, and periodically modulating the effective index of the channel waveguide along the propagation direction. This results in a periodic quantum wire akin to a 1D PhC. Thereafter, interesting properties may be engineered by once again, introducing defects in the periodic structure to access forbidden modes or slightly changing the periodicity to alter the effective bandgap and dispersive properties. The strong confinement of fields in these engineered quantum wire-like structures will also enable interaction with other overlapping fields or discrete components despite the highly guided nature of these modes.

Silicon-on-insulator materials (SOI) and III-V compound semiconductor materials will be used for most of the discussion in this chapter because SOI is compatible with the well established CMOS fabrication process, and III-V semiconductors have been frequently used in heterogeneous integrated circuits and systems. In addition, the large index difference between silicon and its oxide leads to highly confined modes and enables the miniaturization of on-chip silicon-based photonic circuits. Furthermore, silicon is optically transparent and has a very low material absorption coefficient around the telecom wavelength $\lambda = 1550$ nm. Waveguiding loss in SOI platforms has a state of the art value of less than 1 dB/cm. In terms of the impact for future systems applications, it is evident that next generation computing would benefit greatly from all-optical data transfer and processing on a chip. Electrical interconnections inherent in today's computing cannot measure up in terms of both speed and bandwidth. Researchers in the field are aware of the need to bypass any sort of electro-optic process in order to take computing speeds to the next level. Much work is being done in creating both passive and active devices in SOI. Discrete device components such as filters, modulators, and resonators have been demonstrated. Active devices utilizing Raman gain and hybrid silicon lasers, which achieve gain from a bonded III-V material, have also been demonstrated. The momentum of research in this area is the best evidence that silicon photonics is set to revolutionize the field of computing and communications.

In this chapter, we will divide the analysis of the dielectric metamaterials into two categories, namely those in the deep sub-wavelength and those in the larger, sub-wavelength scale. As we shall see, interesting emergent properties arise when the materials are engineered to sizes smaller than or comparable to the wavelength of light in the said medium. Optical field concentrators, compact sources, polarizers, chromatic dispersers, diffractive structures, and other optical processing devices can now be implemented on-chip using metamaterials wherever natural materials with similar properties either do not exist, or (more frequently) would not be compatible with lithographic fabrication processing. Moreover, there exists an opportunity to develop a new family of optical devices exploiting near-field interactions to a much greater extent than has been possible to date. In summary, there exist an opportunity in using advanced lithography and material composition to "lithographically right and assemble" optical materials and devices with *novel optical functionalities* into circuits and subsystems on a chip.

2. Nanophotonics process

To advance this technology, investigations of nanostructures and their interaction with electromagnetic field are critical. Engineers also need appropriate modeling and design tools, new fabrication recipes, and test instruments capable of characterizing on-chip components. The nanophotonics process will help to establish near-field optical systems science and underlying technologies to advance future integrated information systems.

The design of integrated photonic systems is a challenging task, as it not only involves the accurate solution of electromagnetic equations, but also the need to incorporate the material and quantum physics equations into the analysis of near-field interactions. These studies need to be integrated with device fabrication and characterization to verify device concepts and optimize device designs. In this chapter we discuss examples of SOI metamaterials and devices that can be realized using CMOS-compatible fabrication process that we demonstrated recently in our lab. These examples include birefringent elements that utilize a combination of geometry and material properties to separate light into orthogonal polarizations, graded-index lenses, frequency-selective resonators and Bragg gratings, and metal-dielectric nanostructures that can achieve extremely tight field confinement. Some of these example devices are tested using a near field characterization tool, the heterodyne near-field scanning optical microscope (H-NSOM). This microscope uses a fiber probe tapered to about 200–500 nm diameter with an aperture of about 50–200 nm, brought close enough to the nanostructure under test to pick up its evanescent electromagnetic fields. The idea to improve resolution of optical measurements by bringing a subwavelength aperture close to the object of interest was first introduced by Synge in 1928 and experimentally realized only in 1983 by two independently working research groups: Dieter W. Pohl and co-workers at IBM[30] and Aaron Lewis and co-workers at Cornell University.[31] The efficiency of light transmission T through a small aperture ($d << \lambda$) drops rapidly[31] as the aperture size decreases, $T \sim (d/\lambda)^4$. Thus, for realistic aperture of 100 nm and visible wavelengths the transmission efficiency only reaches 10^{-6}–10^{-8}.[32] Such small transmission coefficients demand the use powerful optical sources (often lasers), efficient detection schemes, and detectors. One example of a suitable detection scheme is heterodyne detection, an interferometric technique that not only improves the detection efficiency but also allows measurement of optical phase.

The concept of heterodyne detection is to mix the signal of interest with a coherent reference beam at a slightly shifted optical frequency. This can be done via a Mach-Zehnder interferometer with one (signal) arm including NSOM and the other (reference) arm providing a frequency-shifted reference. The two fields are added coherently, yielding the desired interference signal oscillating at the heterodyne frequency. The coherent gain of the heterodyne detection significantly improves the sensitivity of the instrument.[33] The system is built from readily available telecom components. This in-fiber realization provides better interferometric stability; polarization-maintaining fibers can also be used for maximizing interference term.

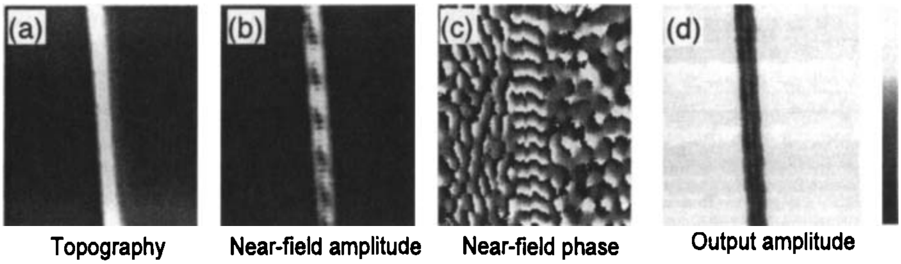


Figure 1. (a) Waveguide topography obtained in NSOM measurements; (b) and (c) near-field amplitude and phase distributions (respectively); and (d) output amplitude *vs.* position of the NSOM tip.

The scanning process provides simultaneously three images: sample topography, deduced from the AFM feedback system that keeps the probe at a constant height above the sample surface; and amplitude and phase distributions of the evanescent optical fields. Example of H-NSOM characterization of an SOI waveguide is shown in Fig. 1. The mapping of evanescent fields has proven to be a powerful tool in understanding the performance of nanoscale optical materials, devices, and circuits.

3. Dielectric metamaterials

In this Section, we will describe the analysis of deep sub-wavelength scale dielectric metamaterials and their behavior in the sub-wavelength scale limit, as well as when they are perturbed into aperiodic composites.

We investigate a class of dielectrics, characterized by feature sizes $\delta << \lambda/2n$, where λ is the free-space wavelength and n is the refractive index of the dielectric material. Photonic structures with periodic or quasi-periodic refractive index variations on a scale much smaller than the wavelength of light can be called "metamaterials" (from the Greek word "μετά", meaning "after" or "beyond") – materials whose optical properties arise from structural geometry rather than only from the composition of constituent materials. This approach, as we will discover in detail in this Section, can be illustrated by the simplest example of form-birefringent materials: 1D periodic structures that have a polarization-dependent index of refraction[34–36] and unusual nonlinear properties.[37,38]

Form birefringence occurs in structures that have deep subwavelength periodicity.[18] The altered surface morphology of the dielectric used to construct such structures results in a large difference between the effective indices of the TE and TM polarized optical fields, since they need to satisfy different boundary conditions. Form-birefringent nanostructures (FBNs) go beyond naturally birefringent materials in (i) the strength of birefringence $\Delta n/n$ (where Δn is the difference in the refractive index for the two orthogonal polarizations); (ii) the extent of form birefringence Δn that may be adjusted by varying the duty cycle as well as the shape of the microstructures; and (iii) the possibility of modifying the reflection properties of dielectric boundaries.[39] These features are useful for constructing polarization-selective beam splitters and general-purpose polarization-selective diffractive optical elements such as birefringent computer-generated holograms.[36] Extending this concept to the 2D geometry or implementing aperiodicity enables other useful functionalities, such as converting a linear polarization state to radial or azimuthal polarization[39,40] and creating graded-index media.[40,41] It was also shown that metamaterial approach can help to overcome fabrication difficulties and create a Fresnel lens analogue using binary lithographic fabrication with deep subwavelength feature sizes of less than 60 nm.[29]

It is also possible to mold the light flow in the planar configuration by using metamaterials. Bringing the functionality of tabletop optical information processing components to a chip will create compact devices, which can benefit from fast data transfer, small form-factor, parallel processing, and low power consumption. Implementing free-space-like propagation for planar optics means that while the light is confined by index difference in the vertical direction (chip plane), the horizontal beam size is regulated by phenomena similar to the diffraction, reflection and refraction of 3D free-space optics. This can be seen as a direct and more natural transition of the conventional free-space bulk optical components and devices to the chip-scale of photonic integrated circuits.

To create a dielectric metamaterial we use a subwavelength structure that can be fabricated in the high refractive index slab – see Fig. 2(a). The slab has an index of refraction n_1, while the gaps in the etched subwavelength structure can be filled with a material possessing a lower index of refraction n_2, *e.g.* air with $n_2 = 1$. This slab structure is constructed on the cladding with an overall lower index of refraction $n_C < n_1$, to ensure confinement in the vertical direction. For some material systems, such as SOI, the cladding with the guiding slab is located on top of a thick substrate with the refractive index n_S. Consider a grating with a period Λ, where $F\Lambda$ is the fraction of the unit cell filled with high-index material. It can be shown that the second-order effective medium theory approximation[42,43] is accurate for small grating periods $\Lambda < \lambda/n$[18] and for grating thickness larger than $\lambda/3$.[44] Other approaches in design and analysis of these subwavelength grating metamaterial structures include numerical methods such as RCWA, finite element method (FEM), and the finite-difference time-domain (FDTD) approaches.

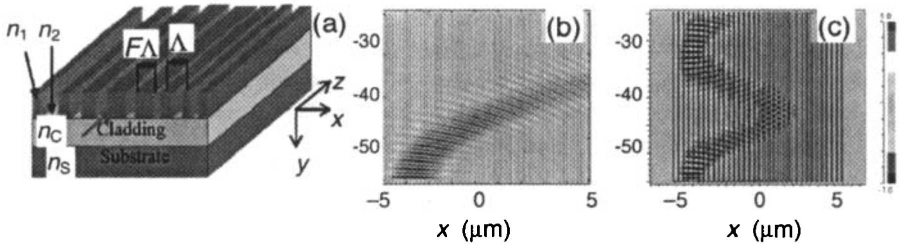


Figure 2. Schematic diagram of a subwavelength planar metamaterial: (a) description of a deeply subwavelength approach to dielectric metamaterial structure; Numeric simulation results showing light propagation in subwavelength SOI gratings with spatial chirp introduced by linearly varying the filling factor (increasing from left to right) and initial periods of (b) $\Lambda = 150$ nm and (c) $\Lambda = 300$ nm.

This concept can be used to create new materials with refractive indices different from those of the constituent materials. For example, for the SOI material system, we usually have silicon with index of refraction of $n_{Si} = 3.48$ and silicon dioxide with $n_{SiO2} = 1.46$ as the only materials available for structure design. On the other hand, in the tabletop free-space optics space we have access to a variety of other materials: different glasses, crystals, polymers, *etc.* This fact makes it difficult to directly translate tabletop optical setups to on-chip implementations. Metamaterials can provide a solution to overcome this difficulty. For example, by fabricating a subwavelength grating, the achievable index in SOI varies from 1.5 to 3.4, thus covering almost fully the range between high-index silicon and low-index oxide. This range was estimated for TE-polarized fields in structures with a period of $\Lambda = 400$ nm satisfying the $\Lambda < \lambda/n = 1500$ nm/3.5 ~ 400 nm condition, with filling factors varying from 0.1 to 0.9 to comply with the state-of-the-art nanofabrication capabilities (feature size ~ 40 nm).

Next we examine a subwavelength structure with a linearly varying filling factor. Such a slab will equivalently act as a graded-index metamaterial, where the effective index of refraction in the transverse direction decreases or increases linearly. It is well known that in such a "graded" index material, the incident beam of light will bends towards the higher index of refraction. We performed numerical simulations of light propagation in such a spatially chirped subwavelength grating structure with the initial periods $\Lambda = 150$ and 300 nm. The numeric simulation results for SOI material platform are summarized in Fig. 2(b) and (c). The propagation of light for the structure the shorter initial period (deep sub-wavelength limit at $\Lambda = 150$ nm) shows truly graded index behavior – see Fig. 2(b). Conversely, for the larger initial period $\Lambda = 300$ nm we can observe some reflection – see Fig. 2(c) – with a "snake-like" propagation trajectory due to Bragg reflection. That is, as the light beam propagates close to the normal to the refractive index gradient, its effective wavelength becomes smaller (as the index increases) until the Bragg matching condition is satisfied when $\lambda_{eff} = \lambda/n = 2\Lambda\sin\theta$ (for the first diffraction order). After the Bragg-matched reflection, the light beam propagates towards the lower effective refractive index in the graded medium. Consequently, the "total internal reflection" condition will be satisfied and the light beam will be reflected back again. This type of "snake-like" light propagation occurring due to a combination of nonresonant and resonant transient metamaterial behavior is quite unusual and cannot be observed in natural materials.

An example of using such a nonresonant metamaterial that can dramatically affect the dispersive properties of the propagating light[45] for a specific application is the mode-matching device, where position-dependent polarizability to guide and manipulate the modes of light propagating on a chip. The device is realized by lithographically defining and etching subwavelength features into a high-refractive-index slab waveguide,[29] modifying its local effective index of refraction. We use a subwavelength periodic structure and locally modulate its duty cycle in the transverse direction x, to achieve modulation of the index of refraction, i.e. $n(x) = n_0(1 - \alpha x^2/2)$, where n_0 and α are constants representing the maximal effective index and the gradient strength, respectively. To validate our approach, we designed and fabricated a graded index slab element that focuses light into a 2 μm wide Si ridge waveguide. We used an SOI wafer with a Si slab thickness of 250 nm and an oxide thickness of 3 μm. The fabricated element in Fig. 3 shows the layout of the device. A grating period of 400 nm was chosen to assure the validity of the effective medium approximation on one hand, and to avoid the need for fabricating ultra small features. For compatibility with CMOS fabrication, the minimal air gap was chosen to be 100 nm, imposing a maximal duty cycle of 75%. An SEM micrograph showing the layout of the entire fabricated device is depicted in Fig. 3.

Typically, characterization of nanophotonic devices is performed by analyzing the light intensity at the output of the device. Unfortunately, this approach lacks the ability to probe the amplitude and, even more importantly, the phase profile of the optical beam as it propagates within the structure. To overcome this problem, the fabricated samples were characterized with the H-NSOM,[33] capable of measuring both amplitude and phase of the propagating optical field with a resolution of about 100 nm. Figure 4 shows the measured amplitude and phase of the optical field propagating through the device at a wavelength of $\lambda = 1550$ nm. Figure 4(a) shows the amplitude of the optical field in the region that includes the input waveguide, the non-patterned slab ("S") and large portion of the slab lens section ("L"). The dashed vertical white lines mark the boundaries between the various sections of the device. Light propagates from left to right. Figure 4(b) shows the measured phase in the same region. Figure 4(c) shows several cross-sections of the phase front calculated from Fig. 4(b) at several planes along the *z*-axis. The obtained results clearly show the expansion of the optical beam in the region of the slab. As expected, the phase front is diverging in this section. As the beam enters the metamaterial, the curvature of the phase front gradually decreases and becomes planar after about 5 μm propagation in the slab. Then, the phase front begins to converge towards the focus. As the beam continues to propagate, the phase front starts to diverge again, and the optical beam expands.

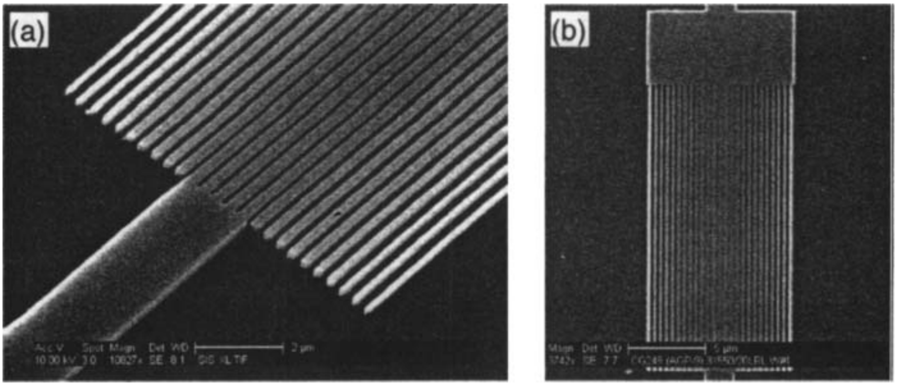


Figure 3. Scanning electron micrograph showing the fabricated device: (a) top view of the entire structure; (b) magnified slanted view showing part of slab lens and the output waveguide.[29]

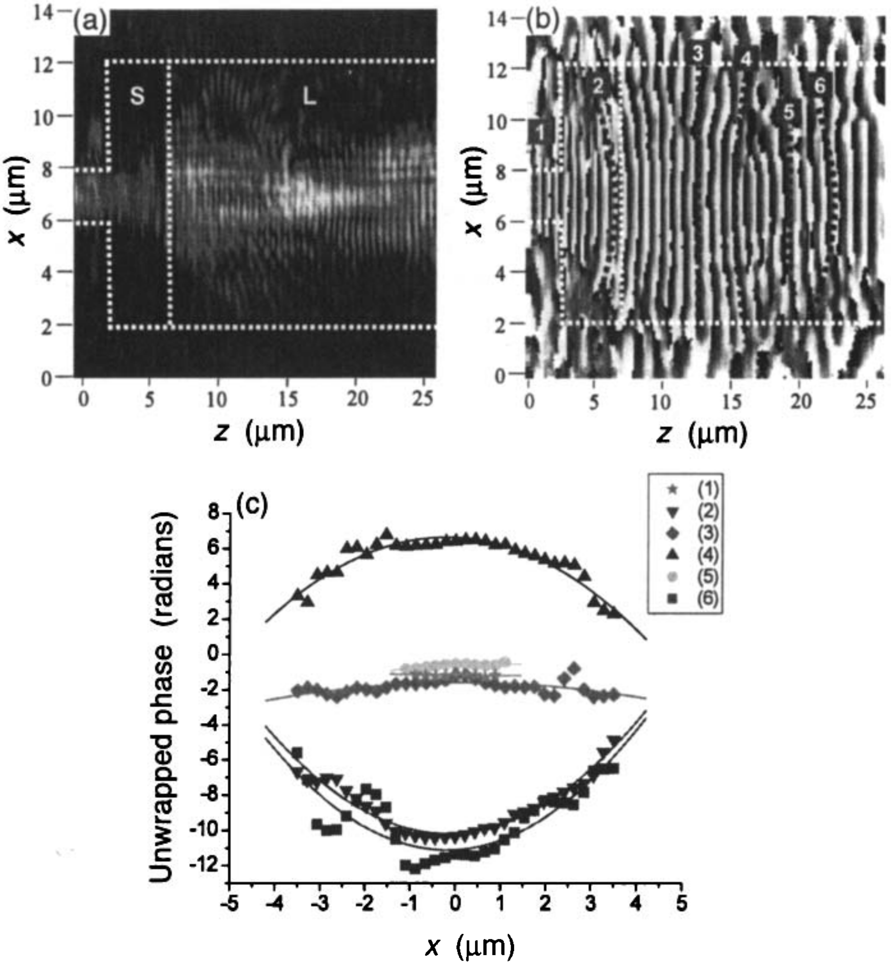


Figure 4. Experimental results obtained with the H-NSOM: (a) amplitude and (b) phase of the optical field in the region that includes the input waveguide, the non-patterned slab ("S") and large portion of the slab lens section ("L"). The dashed vertical white lines mark the boundaries between the various sections and light is propagating from left to right. (c) Cross sections showing phase profile at several planes along the device marked in (b).[29]

The investigated metamaterial-based graded-index slab "lens" device is the first step towards the realization of the FSOC concept. Our H-NSOM measurements clearly demonstrate the focusing effect. This experimental demonstration opens new possibilities in the field of on-chip photonic integration, as the demonstrated component can be integrated with other building blocks (some yet to be developed) to create future devices and systems based on the FSOC concept. We believe that this new concept may become essential for applications such as optical interconnections, information processing, spectroscopy and sensing on a chip.

4. Quantum wires: Subwavelength inhomogeneous dielectrics

We refer to sub-wavelength structures as structures with features that are smaller than but comparable to the wavelength of light. An example of such periodic structures is a general family of resonant structured materials such as photonic crystal (PhC) lattices and the whole family of devices that can be implemented in a PhC lattice. In this section we explore a novel practical approach that we call quantum wire metamaterials that can be used to implement both subwavelength and deep subwavelength nanostructures. We recall that subwavelength scale devices are characterized by feature sizes $\Lambda < \lambda/2n_{\mathrm{eff}}$, where λ is the vacuum wavelength and n_{eff} is the effective index of the specific mode in the device: $n_{\mathrm{eff}} = \beta/k$, where β is the propagation constant of the waveguide mode and k is the wave number in vacuum. For example, for the SOI material system, we can construct a typical single-mode silicon waveguide with $n_{\mathrm{eff}} \sim 2.5$ for operation at the telecommunications wavelength of 1.55 μm. In the following, we will mainly focus on the subwavelength regime to demonstrate the unique capabilities of this approach and demonstrate experimental device prototypes. We first investigate a periodic 1D PhC quantum wire, and extend the investigation to a quasi-periodic quantum wire. Finally, we study the characteristics of a filter created by coupling two such quantum wires together.

The PhC lattice[13–15] is a well known resonant inhomogeneous material, which has been fabricated in many ways during the past decade.[19,20,35,46] It may play a unique role as an integration platform of nanophotonic devices, as numerous PhC-based devices with various functionalities have been reported, including modulators, detectors, filters, lasers, superprisms and elements with negative refraction. However, fabrication of 3D resonant inhomogeneous materials and devices remains challenging, and frequently, 1D and 2D topologies are used due ease in their design, fabrication and integration. A simple example of a 1D resonant structure is a distributed Bragg reflector (DBR) used in planar waveguide technology to perform and enhance various functionalities of optical elements, such as a single-mode selector in semiconductor lasers, optical filters, switches, modulators, couplers, detectors, and sensors. The DBRs conventionally fabricated on the surface of the waveguide involve an additional fabrication procedure separate from the waveguide. Recently, in contrast to traditional approaches, we exploited a single-step fabrication method to define the DBRs and other nanostructured resonant devices using the corrugation of waveguide sidewalls. In this approach, both the period and the modulation strength of the DBRs are lithographically assigned on the waveguide sidewalls.[47–49]

The designed devices are Fabry-Perot (FP) type filters made of a pair of identical Bragg reflectors each having reflection (r) and transmission (t) amplitude coefficients, and separated by a spacer of length $d = \Lambda/2$ causing a phase shift $\phi = \pi/2$. The transmission amplitude of the resonant filter t_{RF} as a function of λ is given by $t_{\mathrm{RF}} = t^2\exp[i\phi]/(1 - r^2\exp[i2\phi])$. These filters are fabricated on a piece (~1x0.5 cm^2) of 6" SOI wafer consisting of a silicon top layer with a mean thickness of 252.2 nm with distribution of $6\sigma \sim 18.3$ nm on a ~3 μm thick silicon dioxide layer.

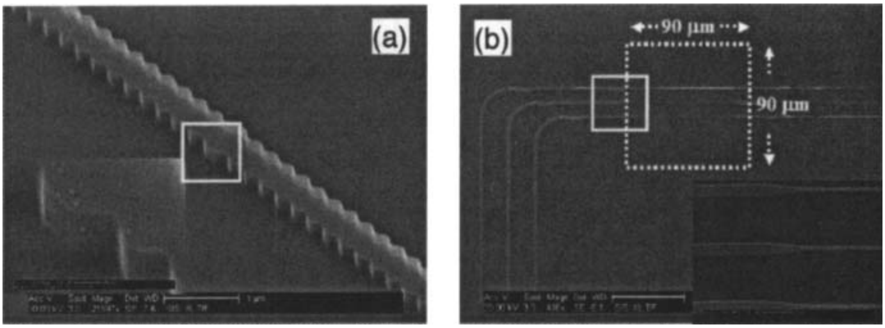


Figure 5. SEM micrographs of the devices. Insets of each figure show the magnified view of the rectangles: (a) transmission resonant filter with vertical gratings; (b) micrograph of the whole device and its inset showing the waveguide connection defined at two different magnifications. Dotted rectangle shows the area defined at x1000 magnification.[47,48]

A tilted SEM image of the fabricated vertical grating filter is shown in Fig. 5(a). Figure 5(b) shows the micrograph of the entire device, whereas the inset of the Fig. 5 shows the magnified view of the tapered waveguide connections. The device is tested by coupling a linearly polarized tunable laser source into a polarization maintaining (PM) tapered fiber with an output spot diameter of ~2.5 μm, producing ~20 dB polarization extinction. Another PM tapered fiber was used to collect light from the fabricated devices and its relative power transmission over the wavelength is analyzed. All measurements were performed using TE polarization. Figure 6 shows the measured transmission spectra of a fabricated device with a wide stopband of ~19 nm bandwidth containing a narrow transmission band of linewidth $\Delta\lambda \sim 0.5$ nm in the center of the stopband. The developed fabrication procedures show very good surface quality, which is also indicated by the measured value of cavity $Q \sim \lambda_B/\Delta\lambda$ of about 3000.

The described resonant PhC-based devices and components implemented using quantum wires are essential for future integration of various information systems on a chip. The introduced graded index device is the first step towards the realization of devices based on near-field optical interactions using space variant polarizability. Moreover, a resonant nanophotonic device utilizing vertical gratings has been realized on an SOI wafer via a fabrication procedure that maintains good surface quality. These experimental demonstrations open new possibilities in the field of on-chip integrated photonic devices, as the demonstrated component can be combined with other building blocks to create future devices and systems based on the concept of free-space optics on a chip. Finally, the use of lithography for sidewall modulation is opening numerous opportunities in constructing various resonant devices for detection, modulation, generation, and manipulation of light on a chip.

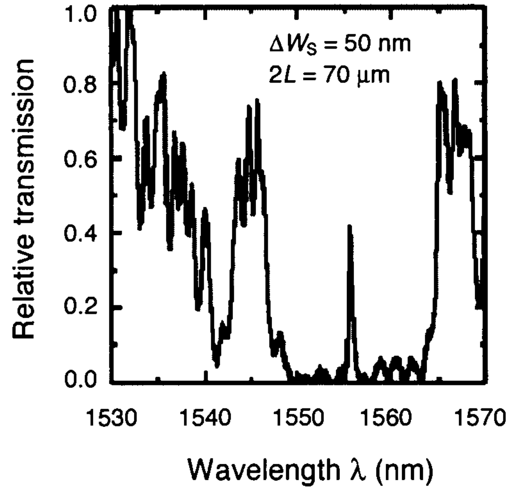


Figure 6. Measured spectra from the fabricated devices with W_S = 500 nm and Λ = 305 nm for ΔW_S = 50 nm, $2L$ = 70 μm.[47, 48]

5. Nanophotonic devices and circuits: Wavelength selective add drop filter with vertical gratings on silicon chip

The SOI material platform is of great interest for information systems integration because it is compatible with the well-established CMOS fabrication process. Additionally, III-V semiconductors have been frequently used in heterogeneous integrated circuits and systems. Furthermore, silicon is optically transparent and has a very low material absorption around the telecommunications wavelength of 1.55 μm. Waveguiding loss in SOI platforms has a state-of-the-art value of less than 1 dB/cm. In terms of the impact for future systems applications, it is evident that next generation computing would benefit greatly from all-optical data transfer and processing on a chip. Electrical interconnections inherent in today's computing cannot measure up in terms of both speed and bandwidth. Researchers in the field are aware of the need to bypass all electro-optic processes in order to take computing speeds to the next level.[50] Much work is being done in creating both passive and active devices in SOI. Discrete device components, such as filters, modulators, and resonators have been demonstrated. For example, the device described in Fig. 5 can serve a near-universal platform, but it somewhat suffers from the limited spectral bandwidth of the stopband of the sidewall-modulated quantum wire which is limited by the achievable effective index modulation. To achieve a larger stopband we need to increase the modulation, but this would cause increased scattering losses that will reduce the quality of the resonator. In this Section, we describe an alternative approach based on a wavelength-selective coupler (WSC) with vertical gratings on silicon chip.[51]

Figure 7(a) shows a schematic diagram of the WSC. We design the structure with three Bragg conditions:[52] $2\beta_1(\lambda) = 2\pi/\Lambda$, producing backward coupling in waveguide 1 (WG1); $2\beta_2(\lambda) = 2\pi/\Lambda$, producing backward coupling in WG2; and $\beta_1(\lambda) + \beta_2(\lambda) = 2\pi/\Lambda$ (cross coupling), where β_1 and β_2 are the propagation constants of WG1 and WG2 respectively, and Λ is the period of the grating. Here, we allocate the cross-coupling wavelength to $\lambda_B = 1.55$ μm as a desired channel wavelength by setting the period as $\Lambda_B = 2\pi/[\beta_1(\lambda_B) + \beta_2(\lambda_B)]$, and backward coupling wavelengths in WG1 and WG2 to outside of C-band by setting $2\beta_1(\lambda > 1.57\ \mu m) = 2\pi/\Lambda_B$ and $2\beta_2(\lambda < 1.53\ \mu m) = 2\pi/\Lambda_B$. Figure 7(b) shows the plots of $2\beta_1$, $2\beta_2$ and $(\beta_1 + \beta_2)$ *vs.* wavelength (calculated with the effective index method), together with transmission spectrum from two-dimensional FDTD method (input at Port 1). The structure dimensions are described in Fig. 7(a) and all the calculations here are performed for the quasi-TE mode. We can see that in this design, we obtain the cross coupling with around 9 dB extinction at 1.55 μm and the backward couplings outside the C-band at 1.52 μm and 1.59 μm.

We fabricated the WSC designed above using e-beam writing and reactive ion etching processes on an SOI wafer with 3 μm buried oxide layer and 250 nm silicon layer. The SEM micrographs of the fabricated structure are shown in Fig. 8. After depositing the over-cladding SiO_2 layer of 2 μm thickness, we measured the transmission spectrum using a broadband light source input at Port 1 and a spectrum analyzer (see the inset of Fig. 8). We observe that the measured spectrum at Port 2 and 3 are in good agreement with the simulated spectrum. In order to further improve the extinction, we simply increase the total length or the waveguide modulation, or decrease the gap between the waveguides. And also, we could extend this device to multi-port add-drop device with placing another waveguide on the other side of the WG1 and/or cascading the WSCs.

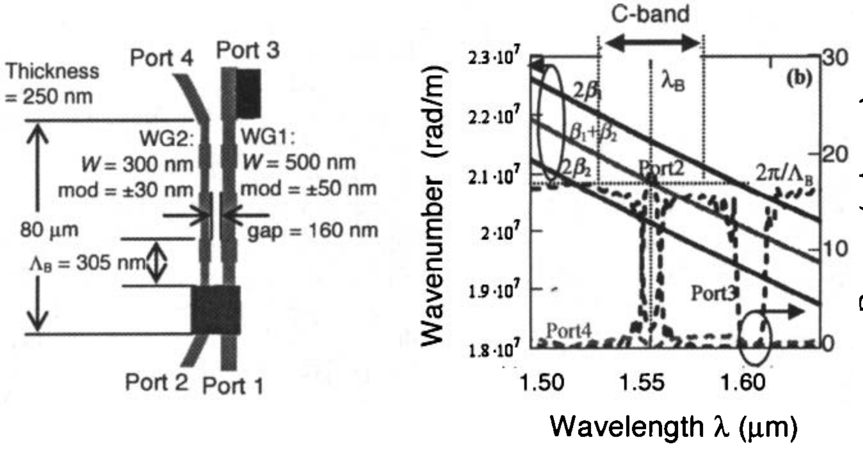


Figure 7. (a) Schematic diagram and detail dimensions of WSC with vertical gratings; (b) plots of three Bragg conditions and transmission spectrum from 2D FDTD.[51]

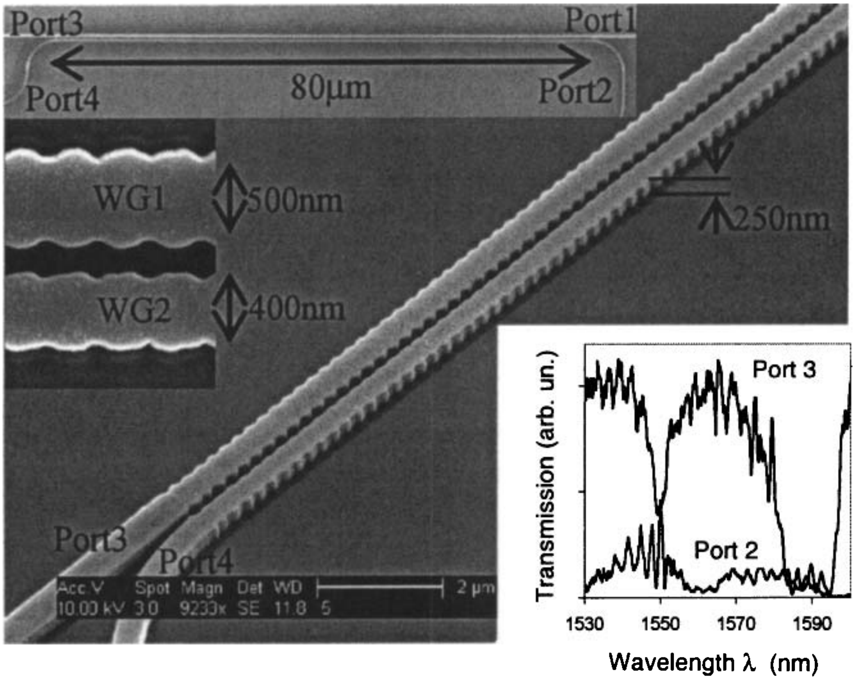


Figure 8. SEM micrographs of the fabricated WSC and measured transmission spectrum at Port 2 and 3.[51]

6. Discussion and future perspectives

This chapter has given an introductory view on the emerging field of nanophotonics and its potential for chip-scale integration of information systems. The main points include our ability to use standard and emerging nanofabrication tools to create complex geometries of composite materials varying the spatial distributions periodically, quasi-periodically or even at random. The feature sizes of these nanocomposites can be deeply subwavelength, which enables creation of materials and devices with unique functionalities that are impossible to achieve with existing technologies. Moreover, the subwavelength structures can incorporate coupled devices operated in spatial as well as spectral resonance. We have described the localized effects and alterations in optical behavior of dielectric metamaterials with space-variant polarizability (form birefringence, snake-like propagation, and metamaterial lensing), as well as various functionalities achieved in single and spatially-coupled resonant quantum wires.

The current trend of using the SOI material platform will continue due to its compatibility with well-established CMOS fabrication, which is projected to reach resolution on the order of 10–16 nm in the near future. Another major reason is that future computing systems complexity will increasingly rely on optical interconnections, possibly including intra-chip communications, where silicon photonics will enable efficient manipulation, transformation, filtering and detection of light meeting the needs of the future information systems. It should be noted that efficient generation of light on a silicon chip is still in its infancy and may not be able to overcome the fundamental problem of the indirect bandgap, but alternative solutions similar to delivery of electrical power from off–chip sources will bring optical fields into Si chips. Progress in heterogeneous metamaterials is certainly another valid alternative, but it may require major technological and manufacturing breakthroughs. It is evident that the fundamental limits of scale and composition in the present technology will continue to shift to ever smaller features and ever more complex material compositions. It is possible that current trends in engineering electronic bandgaps are not merely a pipe dream, and may one day merge with engineered nanophotonic metamaterials, leading to new device concepts that will serve as the backbone of future information systems technologies. Engineering the polarizability of dielectric nanostructures is still in its nascent stages and numerous novel functionalities are still to come.

References

1. Y. Fainman, E. Klancnik and S. H. Lee, "Optimal coherent image amplification by two beam coupling in photorefractive $BaTiO_3$," *Opt. Eng.* **25**, 228 (1986).
2. Y. Fainman, C. C. Guest, and S. H. Lee, "Optical digital logic operations by two beam coupling in photorefractive material," *Appl. Opt.* **25**, 1598 (1986).
3. P. Ambs, S. H. Lee, Q. Tian, and Y. Fainman, "Optical implementation of the Hough transform by a matrix of holograms," *Appl. Opt.* **25**, 4039 (1986).
4. P. Ambs, Y. Fainman, S. H. Lee, and J. Gresser, "Computerized design and generation of space-variant holographic filters. Part I: System design considerations and applications of space-variant filters to image processing," *Appl. Opt.* **27**, 4753 (1988).
5. Q. Tian, Y. Fainman, Z. H. Gu, and S. H. Lee, "Comparison of statistical pattern recognition algorithms for hybrid processing. Part I: Linear mapping algorithms," *J. Opt. Soc. Amer. A* **5**, 1655 (1988).
6. Q. Tian, Y. Fainman, and S. H. Lee: "Comparison of statistical pattern recognition algorithms for hybrid processing. Part II: Eigenvector-based algorithms," *J. Opt. Soc. Amer. A* **5**, 1670 (1988).
7. J. Y. Jau, Y. Fainman, and S. H. Lee: "Comparison of artificial neural network with pattern recognition and signal processing," *Appl. Opt.* **28**, 302 (1989).
8. H. Rajbenbach, Y. Fainman and S. H. Lee: "Optical implementation of an iterative algorithm for matrix inversion," *Appl. Opt.* **26**, 1024 (1987).
9. P. C. Sun, Y. Mazurenko, W. Chang, P. Yu, and Y. Fainman " All-optical parallel-to-serial conversion by holographic spatial-to-temporal frequency encoding," *Opt. Lett.* **20**, 1728 (1995).

10. P. C. Sun, Y. Mazurenko, and Y. Fainman, "Femtosecond pulse imaging: Ultrafast optical oscilloscope," *J. Opt. Soc. Amer. A* **14**, 1159 (1997).
11. D. M. Marom, D. Panasenko, P. C. Sun, and Y. Fainman, "Spatial-temporal wave mixing for space-time conversion," *Opt. Lett.* **24**, 563 (1999).
12. R. E. Saperstein, D. Panasenko, and Y. Fainman, "Demonstration of a microwave spectrum analyzer based on time-domain optical processing in fiber," *Opt. Lett.* **29**, 501 (2004).
13. J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade, *Photonic Crystals: Molding the Flow of Light*, 2nd ed., Princeton, NJ: Princeton University Press, 2008.
14. E. Yablonovitch, "Inhibited spontaneous emission in solid-state physics and electronics," *Phys. Rev. Lett.* **58**, 2059 (1987).
15. S. John, "Strong localization of photons in certain disordered dielectric superlattices," *Phys. Rev. Lett.* **58**, 2486 (1987).
16. M. Abashin, P. Tortora, I. Marki, *et al.*, "Near-field characterization of propagating optical modes in photonic crystal waveguides," *Opt. Express* **14**, 1643 (2006).
17. R. D. Meade, A. Devenyi, J. D. Joannopoulos, O. L. Alerhand, D. A. Smith, and K. Kash, "Novel applications of photonic band gap materials: Low-loss bends and high Q cavities," *J. Appl. Phys.* **75**, 4753 (1994).
18. I. Richter, P. C. Sun, F. Xu, and Y. Fainman, "Design considerations of form birefringent microstructures," *Appl. Opt.* **34**, 2421 (1995).
19. R. Tyan, P. C. Sun, A. Scherer, and Y. Fainman, "Polarizing beam splitter based on the anosotropic spectral reflectivity characteristic of form-birefringent multilayer grating," *Opt. Lett.* **21**, 761 (1996).
20. W. Nakagawa, P. C. Sun, C. H. Chen, and Y. Fainman, "Wide field of view narrow band spectral filter base on photonic crystal nanocavities," *Opt. Lett.* **27**, 191 (2002).
21. S. Lin, E. Chow, V. Hietala, P. R. Villeneuve, and J. D. Joannopoulos, "Experimental demonstration of guiding and bending of electromagnetic waves in a photonic crystal," *Science* **282**, 274 (1998).
22. H. Kosaka, T. Kawashima, A. Tomita, M. Notomi, T. Tamamura, T. Sato, and S. Kawakami, "Self-collimating phenomena in photonic crystals," *Appl. Phys. Lett.* **74**, 1212 (1999).
23. E. Schonbrun, Q. Wu, W. Park, T. Yamashita, C. J. Summers, M. Abashin, and Y. Fainman, "Wave front evolution of negatively refracted waves in a photonic crystal," *Appl. Phys. Lett.* **90**, 041113 (2007)
24. H. Kosaka, T. Kawashima, A. Tomita, M. Notomi, T. Tamamura, T. Sato, and S. Kawakami, "Superprism phenomena in photonic crystals: Towards microscale lightwave circuits," *Phys. Rev. B* **58**, R10096 (1998).
25. E. Schonbrun, M. Abashin, J. Blair, Q. Wu, W. Park, Y. Fainman, and C. J. Summers, "Total internal reflection photonic crystal prism," *Opt. Express* **15**, 8065 (2007).
26. C. Luo, S. G. Johnson, and J. D. Joannopoulos, "Subwavelength imaging in photonic crystals," *Phys, Rev. B* **68**, 045115 (2003).

27. Y. Fink, J. N. Winn, S. Fan, C. Chen, J. Michel, J. D. Joannopoulos, and E. L. Thomas, "A dielectric omnidirectional reflector," *Science* **282**, 1679 (1998).
28. O. Painter, R. K. Lee, A. Scherer, A. Yariv, J. D. O'Brien, P. D. Dapkus, and I. Kim, "Two-dimensional photonic band-gap defect mode laser," *Science* **284**, 1819 (1999).
29. U. Levy, M. Abashin, K. Ikeda, A. Krishnamoorthy, J. Cunningham, and Y. Fainman, "Inhomogeneous dielectric metamaterials with space-variant polarizability," *Phys. Rev. Lett.* **98**, 243901 (2007).
30. D. W. Pohl, W. Denk, and M. Lanz, "Optical stethoscopy: image recording with resolution $\lambda/20$," *Appl. Phys. Lett.* **44**, 651 (1984).
31. A. Lewis, M. Isaacson, A. Harootunian, *et al.*, "Development of a 500 Å spatial resolution light microscope. I. Light is efficiently transmitted through $\lambda/16$ diameter apertures," *Ultramicroscopy* **13**, 227 (1984).
32. G. A. Valaskovic, M. Holton, and G. H. Morrison, "Parameter control, characterization, and optimization in the fabrication of optical fiber near-field probes," *Appl. Opt.* **34**, 1215 (1995).
33. A. Nesci and Y. Fainman, "Complex amplitude of an ultrashort pulse with femtosecond resolution in a waveguide using a coherent NSOM at 1550 nm," *Proc. SPIE* **5181**, 62 (2003).
34. F. Xu, R. Tyan, P. C. Sun, C. Cheng, A. Scherer, and Y. Fainman, "Fabrication, modeling, and characterization of form-birefringent nanostructures," *Opt. Lett.* **20**, 2457 (1995).
35. R.-C. Tyan, A. A. Salvekar, H.-P. Chou, *et al.*, "Design, fabrication, and characterization of form-birefringent multilayer polarizing beam splitter," *J. Opt. Soc. Amer. A* **14**, 1627 (1997).
36. F. Xu, R.-C. Tyan, P.-C. Sun, Y. Fainman, C.-C. Cheng, and A. Scherer, "Form-birefringent computer-generated holograms," *Opt. Lett.* **21**, 1513 (1996).
37. W. Nakagawa, R.-C. Tyan, and Y. Fainman, "Analysis of enhanced second-harmonic generation in periodic nanostructures using modified rigorous coupled-wave analysis in the undepleted-pump approximation," *J. Opt. Soc. Amer. A* **19**, 1919 (2002).
38. W. Nakagawa, R. Tyan, and Y. Fainman, "Analysis of enhanced second-harmonic generation in periodic nanostructures using modified rigorous coupled-wave analysis in the undepleted-pump approximation," *J. Opt. Soc. Amer. A* **19**, 1919–1928 (2002)
39. U. Levy, C. H. Tsai, L. Pang, and Y. Fainman, "Engineering space-variant inhomogeneous media for polarization control," Opt. Lett. 29, 1718–1720 (2004).
40. U. Levy, M. Nezhad, H.-C. Kim, C.-H. Tsai, L. Pang, and Y. Fainman, "Implementation of a graded-index medium by use of subwavelength structures with graded fill factor," *J. Opt. Soc. Amer. A* **22**, 724 (2005).
41. J. N. Mait, A. Scherer, O. Dial, D. W. Prather, and X. Gao, "Diffractive lens fabricated with binary features less than 60 nm," *Opt. Lett.* **25**, 381 (2000).

42. S. M. Rytov, "Electromagnetic properties of a finely stratified medium," *Sov. Phys. JETP* **2**, 466 (1956).
43. E. N. Glytsis and T. K. Gaylord, "High-spatial-frequency binary and multilevel stairstep gratings: Polarization-selective mirrors and broadband antireflection surfaces," *Appl. Opt.* **31**, 4459 (1992).
44. P. Lalanne and D. L. Lalanne, "Depth dependence of the effective properties of subwavelength gratings," *J. Opt. Soc. Amer. A* **14**, 450 (1997).
45. U. Levy and Y. Fainman, "Dispersion properties of inhomogeneous nanostructures," *J. Opt. Soc. Amer. A* **21**, 881 (2004).
46. C. C. Cheng, A. Scherer, R. C. Tyan, Y. Fainman, G. Witzgall, and E. Yablonovitch, "New fabrication techniques for high quality photonic crystals," *J. Vac. Sci. Technol. B* **15**, 2764 (1997).
47. H. C. Kim, K. Ikeda, and Y. Fainman, "Resonant waveguide device with vertical gratings," *Opt. Lett.* **32**, 539 (2007).
48. H. C. Kim, K. Ikeda, and Y. Fainman, "Tunable transmission resonant filter and modulator with vertical gratings," *J. Lightwave Technol.* **25**, 1147 (2007).
49. D. T. H. Tan, K. Ikeda, R. E. Saperstein, B. Slutsky, and Y. Fainman, "Chip-scale dispersion engineering using chirped vertical gratings," *Opt. Lett.* **33**, 3013 (2008).
50. V. R. Almeida, C. A. Barrios, R. R. Panepucci, and M. Lipson, "All-optical control of light on a silicon chip," *Nature* **431**, 1081 (2004).
51. K. Ikeda, M. Nezhad, and Y. Fainman, "Wavelength selective coupler with vertical gratings on silicon chip," *Appl. Phys. Lett.* **92**, 201111 (2008).
52. P. Yeh and H. F. Taylor, "Contradirectional frequency-selective couplers for guided-wave optics," *Appl. Opt.* **19**, 2848 (1980).

What Will Modern Photonics Contribute to the Development of Future Optical Communication Technology?

Djafar K. Mynbaev
Dept. of Electrical and Telecommunications Engineering Technology,
New York City College of Technology–CUNY, Brooklyn, NY 11201, U.S.A.

1. Introduction

Optical communications is the linchpin of modern telecommunications; in fact, the vast majority of global telecommunications traffic today is delivered optically, which means that optical communications laid the foundation for all the beneficial changes in our lives that the Internet now provides. The optical communications industry, in fact, is one of the largest consumers of electronic and photonic products; hence, developments in the optical communications area will, to a great extent, shape trends in electronics and photonics.

Optical communication today ranges widely in three dimensions: in space, in time, and in frequency (spectrum).

In space (that is, transmission distance), optical communication ranges from nanometers in optical interconnects to tens of thousands of kilometers in intercontinental links. In time, the optical communications industry operates from kilobits per second in short-reach connections to terabits per second in long-distance links; these bit rates correspond to characteristic times that range from milliseconds to picoseconds. In spectrum, optical communication covers the wavelength range from 850 nm (local) to 1650 nm (long-haul).

If this variety of the characteristics is insufficient to stress the complexity of modern optical communications, we must add another factor: the network. Delivering information to an individual receiver through a web of connections ranging from chip-level interconnects to the worldwide network is a herculean task. It involves intelligent firmware systems, advanced transmission technology and complex network management, all of which are controlled by quite sophisticated software.

Optical communication engineers are under constant pressure to improve all aspects of this technology – from the ever-increasing demand for bandwidth to reducing the bit-per-kilometer cost. Responding to these demands, they have developed many wonderful components, devices, and systems. However, today's achievements will not guarantee the fulfillment of tomorrow's tasks.

Optical communication, in its hardware layer, is a part of the photonics community, which is actively conducting fundamental and applied research to discover new phenomena and applications. To name just a few, these discoveries

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

range from today's experiments with single-photon manipulation (generation, modulation, transmission, and detection), to silicon-based photonics, to optically active integrated circuits, to photonic crystals, to quantum routers, to metamaterials and to nanophotonics. Such discoveries will transform the optical communication technology.

The purpose of this chapter is to depict the current status of optical communications, to stress its achievements, and to emphasize its needs. We will also discuss several examples of the latest developments in photonics that would enable the development of new components, devices, and systems for fiber-optic communications technology. The hope is that this presentation will stimulate some thoughts and new ideas that will result in the further development of photonics and optical communications.

2. Optical communications

- *General consideration and the scale of operation*

A basic fiber-optic communications system (link) is shown in Fig. 1(a). In general, the transmitter (Tx) is a device that launches an information signal into the optical fiber, which is the conductor of the optical signal, and the receiver (Rx) is a device that extracts information from the optical fiber in a suitable form. We need to remember that the goal of this system is to deliver information. Figure 1(b)

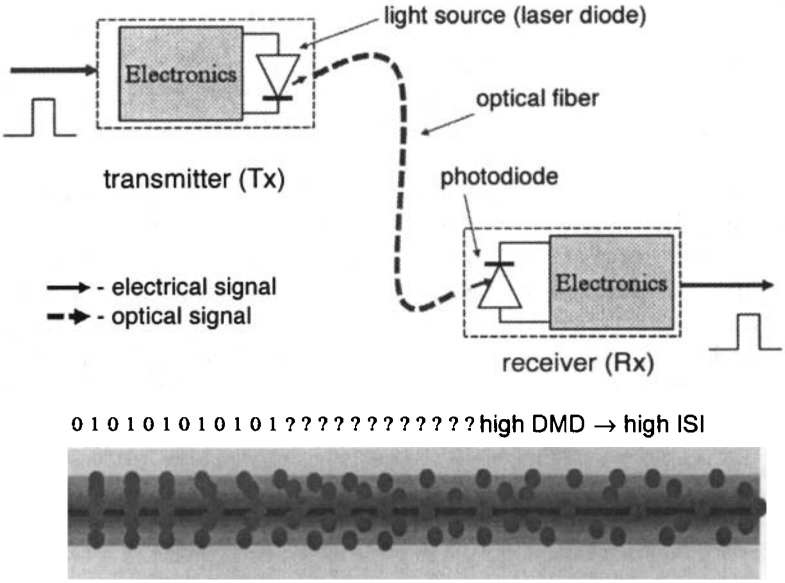


Figure 1. Basics of optical communications: (a) block diagram of a fiber-optic communications link; (b) the goal is to deliver information, not light (DMD – differential mode delay; ISI – inter-symbol interference).

shows the situation when light is delivered but information is lost because of the spreading of optical pulses caused by modal dispersion.[1]

Over the last 30 years, optical communications has developed from a nonexistent technology to the linchpin of the telecommunications industry. Why? The main advantage of optical communications over any other means of transmission is speed. We have to distinguish between two speeds when delivering signals: propagation speed and transmission speed. Propagation speed measured in m/s is the area of interest for developers of on-chip-level optical interconnects; here, since photons move faster than electrons, optical communications offers the solution to interconnect delay. Transmission speed (transmission capacity, or bandwidth) measured in bit/s is proportional to the frequency of a carrier f_C, and since light has the highest f_C (hundreds of THz) of all the practical signal carriers, optical communications systems can achieve the highest transmission capacity and the highest values of bandwidth-distance product.

To better appreciate the scale of optical-communications technology, we need to consider three dimensions of its operation: space (distance), time (bit rate), and spectrum (range of wavelengths).

- *Transmission distances in optical communications*

Let us consider the range of transmission distances covered by optical communications today, as shown in Table 1.[2] Since the cost/performance curve of optics is much steeper than that of electronics, optical communications has effectively replaced the electrical systems in long-distance, metro and access scales.

Figure 2 shows a version of the worldwide optical network map. This map does not show all of the intercontinental, let alone all terrestrial, networks; it does show, however, the global scale of optical communications.

Figure 3 shows an example of a metro-scale optical network developed by just one company, RCN Corporation, to cover New York City. One can only imagine how many similar networks cover this city.

An individual customer can optically access the global telecommunications infrastructure by means of a cable-TV network (CATV) or a passive optical network (PON), as illustrated in Fig. 4.

	World wide networks	Metro scale networks	Local (access) networks	Rack-to-rack links	Card-to-card	On-card links	On-chip links
Distance scale	1000 km	100 km	10 km	10 m	cm	mm–µm	µm–nm
Bit rate in Gb/s per channel	40 Gb/s	10 Gb/s	2.5 Gb/s	up to 2.5 Gb/s	2.5 Gb/s	2.5 Gb/s	2.5 Gb/s
Optics in use	from 1980's	from 1990's	from 1990's	from 2000's	2010's	develop-ment	R/D

Table 1. Optical communications – transmission distances and applications.

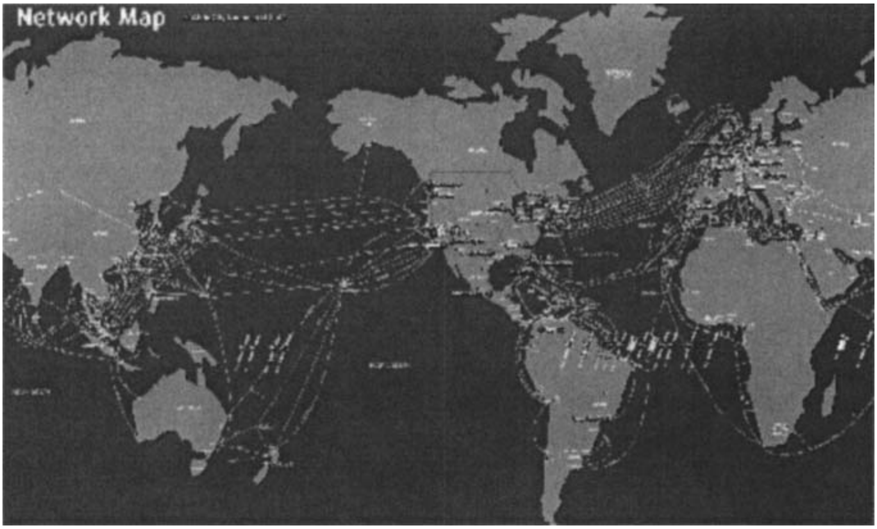


Figure 2. Global optical network (from *http://personalpages.machester.ac.uk/staff/m.dodge/cybergeography/atlas/cables.html*).

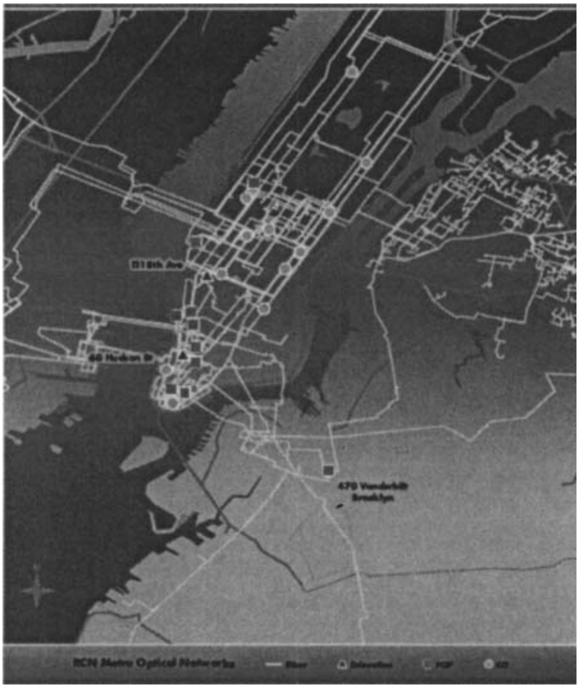


Figure 3. An example of a metro-scale optical network: New York City (courtesy of RCN).

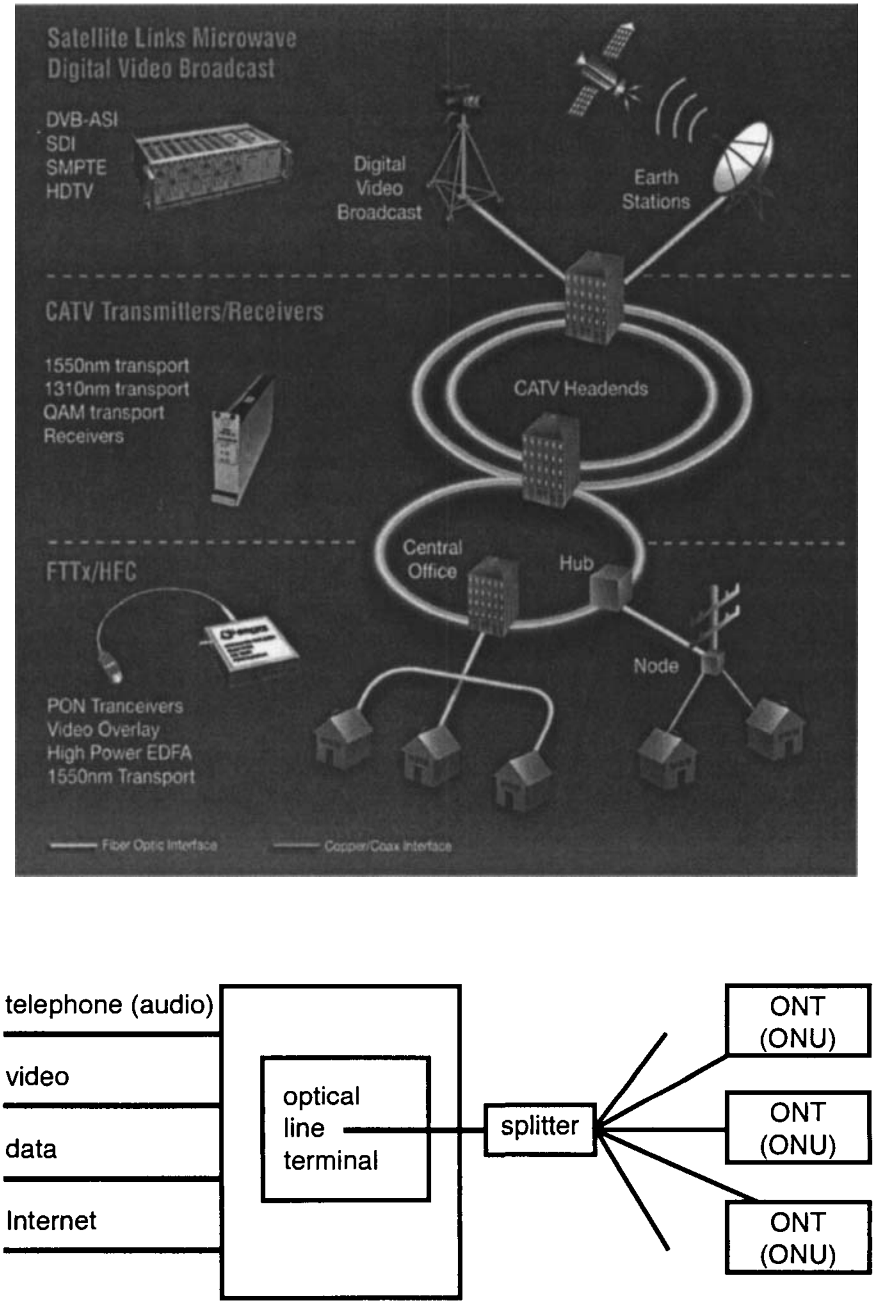


Figure 4. Optical access networks: cable-TV (CATV) network (top); passive optical network (PON) (bottom).

A CATV network is mostly fiber-based; however, there are still segments of coaxial-cable connections. In the United States, a cable-TV network is the primary provider of broadband access, whereas most of the developed world still relies on copper-based DSL technology. A passive optical network (PON) replaces the traditional copper-wire telephone connections from a dwelling to a central office. The installation of PON and other types of optical access networks keeps growing in the U.S. and other developed countries; as a matter of fact, in South Korea and Japan, optical access has become the main type of broadband connection.

Commercially available PONs (*e.g.*, FiOS) are based on time-division multiplexing (TDM). The TDM technique is inherently restricted in bit rate; what is more, TDM-based PON has security and confidentiality issues that still have to be resolved. The natural solution would be the move to wavelength-division multiplexing (WDM). However, the classical WDM approach, as applied to PON,

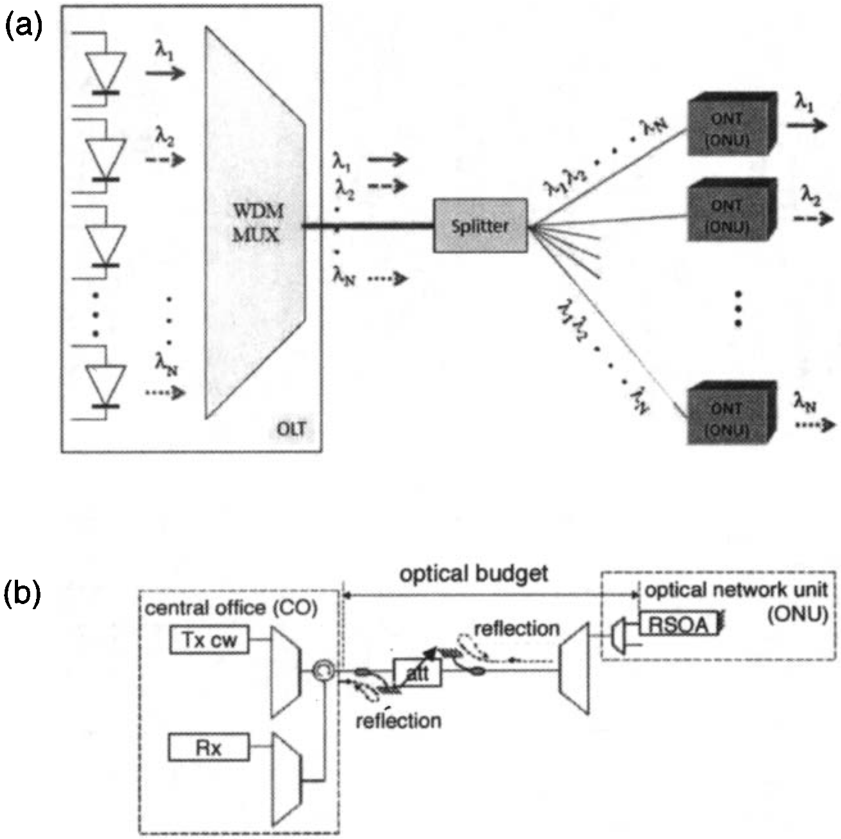


Figure 5. Wavelength-division multiplexing PON: (a) classical approach; (b) colorless implementation.[3]

suffers from a major drawback: it is economically ruinous to keep an inventory of individual optical network terminals (ONTs), also called units (ONUs), for every customer – see Fig. 5(a). Instead, all customers must have the same type of ONT, which requires developing colorless ONTs, as shown in Fig. 5(b).[3]

As we see, optical communications technology replaces copper everywhere, from worldwide to metro to access networks. Also, at the in-room and intra-rack scales, optical fiber has completely replaced copper wire; at the card-to-card and on-card scales, the fiber-optics links start to compete with the electrical links. The trend to replace copper at the chip level is made necessary by continual increases in the bit rate at the system (*e.g.*, computer) and board levels. The movement in computer-system architecture from parallel data buses to high-speed serial I/Os is also in concert with the shift to optics. When the data rate changes from 10 Gb/s to 40 Gb/s and 100 Gb/s, the move to optical interconnects will become inevitable.

The classification of interconnects is based on their applications, which boils down to the interconnect lengths.[2] Thus, we distinguish rack-to-rack (in-room), card-to-card, intra-card, intra-module (on-MCM), and intra-chip (on-chip) interconnects – see Table 1. Today's electrical interconnects suffer from many problems:[4-7] the small value of the bandwidth-length product, significant delay (~20–25 ps/mm), low noise immunity, crosstalk between channels, general noise, high power consumption (~1 mW/mm) accompanied by heat dissipation, and frequency-dependent losses. Also, there are architectural problems with electrical interconnects: the performance of the CPU and other electronic modules increases much faster than the transmission speed of the links among these modules. By some estimates, wire delay will increase 20 times faster than device-level gate delay.

Optical interconnects may be the solution to all these problems. Figure 6(a) shows an interconnect module that includes all main components shown in Fig. 1. Theoretically, optical interconnects will reduce delays (*e.g.*, a delay might be as low as 4–5 ps/mm for a polymer and 10 ps/mm for Si waveguides[8]), will allow for 3D dense packaging thanks to low crosstalk (*e.g.*, the minimum pitch is 8–9 μm for polymer and 0.5–1 μm for Si[8]), and will eliminate or reduce all other drawbacks of electrical interconnects in a similar manner.

There are several areas in the computer industry that are in dire need of optical interconnects. One of these is the area of supercomputers because they rely heavily on transmission between CPUs and memory and on card-to-card (inter-node) transmission. The companies involved in designing supercomputers are actively working on the development of optical interconnects,[5,6] see Fig. 6. Another area where we expect the realization of chip-to-chip optical interconnects is microprocessors, because the microprocessor transition from multi-core to many-core architectures will drive chip-to-chip bandwidth demands. Future architectures will require bandwidths of 200 GB/s to 1.0 TB/s and will bring about the era of tera-scale computing. Currently, a DARPA program named UNIC is funding a five-year project (at Sun Microsystems) to develop microchip interconnectivity via on-chip optical networks based on silicon photonics and proximity communication.[7]

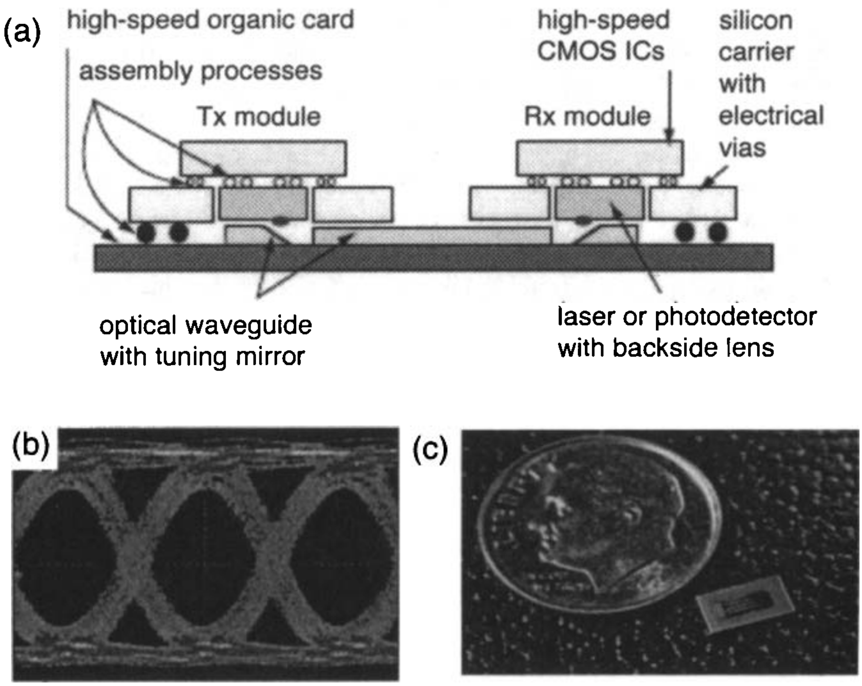

(a)
high-speed organic card
assembly processes
Tx module
high-speed CMOS ICs
silicon carrier with electrical vias
Rx module
optical waveguide with tuning mirror
laser or photodetector with backside lens
(b)
(c)

Figure 6. A Terabus optical interconnect: (a) cross-section of a Terabus module structure; (b) eye diagram at the output of a Terabus receiver with the optical input from a Terabus transmitter at 14 Gb/s; (c) a Terabus chip — an optical transceiver with CMOS driver and receiver chips coupled with III-V optical components, the entire integrated package measures 3.25 by 5.25 mm.[2]

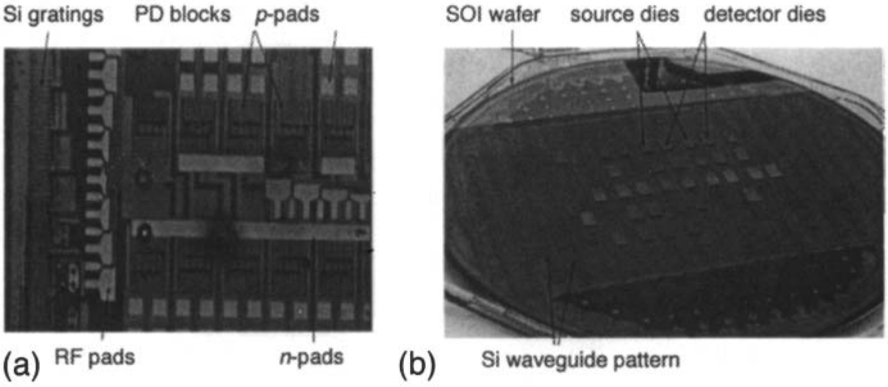


Figure 7. (a) Source and detector dies bonded onto an SOI wafer; (b) a fabricated chip with 10 PD blocks with eight devices per block (courtesy of Pietro Binetti[9]).

Figure 7 demonstrates examples of an optical interconnect fabricated in the research laboratory, their operation, and their dimensions. Bear in mind that card-to-card optical interconnects are commercially available.

To sum up our discussion of the *scale of operation of optical communications*, it is clear that optical communication consistently outperforms its electronic counterpart over ever-shorter distances. For last twenty plus years, optical communications has replaced electronics in scales ranging from ultra-long-haul to rack-to-rack links. These replacements occur mainly in point-to-point topologies; this is why some researchers refer to optical interconnections rather than to networks. In any event, the trend is clear: optical interconnections are increasingly coming closer to inter-chip transmission.

- *Time scaling in optical communications – TDM*

The main advantage of optical fiber over any other transmission medium is its greater bandwidth (transmission capacity). This is why the primary objective in using optical communications has been and remains the increase in the bit rate of the transmitting signal. To discuss this topic, we need to distinguish between the bit rate (BR) of a single channel (at a given wavelength) achieved by time-division multiplexing (TDM) and the data rate (DR) transmitted over a fiber achieved by wavelength-division multiplexing (WDM). The DR for long-distance optical networks runs into tens of Tb/s, whereas metro optical networks operate at tens of Gb/s and access optical networks operate at a few Gb/s or hundreds of Mb/s.

Concerning the TDM BR, a single wavelength now routinely achieves a bit rate of 40 Gb/s. This has been accomplished by improving the performance of individual transmitters and receivers and combining (interleaving) several low-rate signals in one stream (*e.g.*, four 10 Gb/s signals). The industry has been approaching a bit rate of 100 Gb/s.

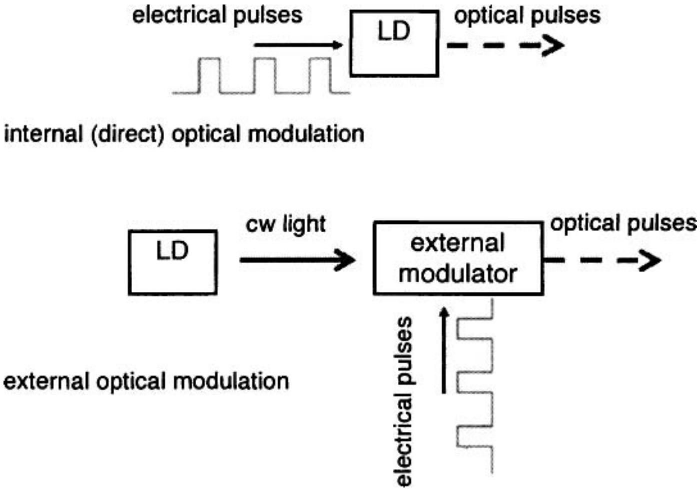


Figure 8. Internal and external modulation techniques.

The TDM bit rate, as well as the transmission performance in general, depends very much on the modulation technique. In optical communications, we employ internal and external modulation schemes, as shown in Fig. 8. Modern laser diodes (LDs) do not allow for internal (direct) modulation at a high bit rate. External modulation keeps LDs in the comfortable cw regime and relegates all modulation problems to an external modulator. With this approach, the rate of 10 Gb/s is routinely achieved.[1] Recently developed modulation schemes, such as quadrature phase-shift keying, overcome dispersion-caused signal degradation without recourse to dispersion-compensation modules, an industry advance that significantly reduces the cost of transmission. The use of an advanced modulation technique is an example of solving a problem with a component (here, the problem lies with the optical fiber) by resorting to a network-level approach.

- *The spectrum of optical communications – WDM*

We need to realize that TDM is inherently restricted in its bit rate. Indeed, for the bit rate of 1 Tb/s, the bit time is equal to 10^{-12} s or 1 ps. There are no photonic devices (LDs, modulators, or PDs) that can operate at this speed. Today's electronics mainly operate at 10 Gb/s; modulation at 40 Gb/s has been achieved and modulation at 100 Gb/s is on the horizon. On the other hand, an optical fiber can carry 100 Tb/s on a single wavelength, which means that an optical fiber can carry a thousand 100 Gb/s channels.

The solution to this bottleneck is WDM, which allows for an n-fold increase in the data rate, where n is the number of wavelengths. An example of four-wavelength WDM and the corresponding frequency spectrum are shown in Fig. 9.

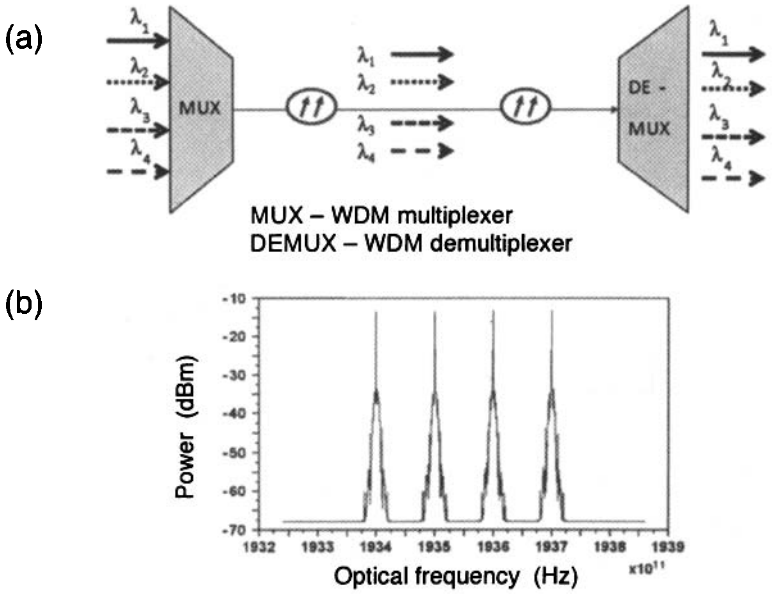


Figure 9. Schematic concept of WDM (a) and an example of a WDM frequency spectrum (b).

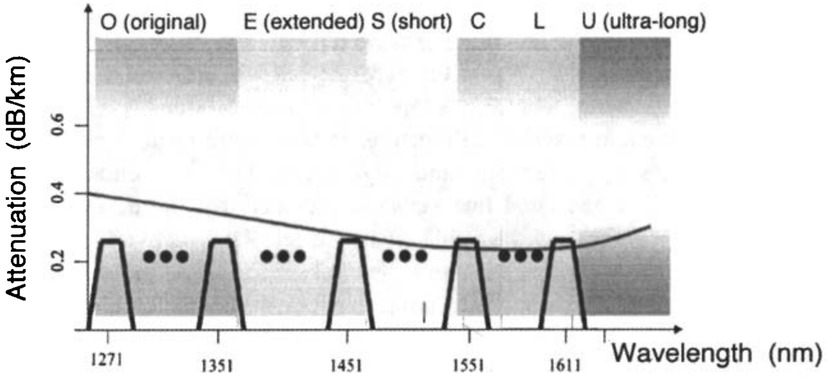


Figure 10. The spectrum and channel spacing of CWDM *vs.* attenuation length (DWDM operates over the C and L bands).

There are two types of WDM today: dense WDM (DWDM) and coarse WDM (CWDM). They differ in spacing between two adjacent channels (wavelengths): DWDM operates at spacing ranging from 1.6 nm (200 GHz) to 0.2 nm (25 GHz); CDWM works at 20 nm spacing. DWDM occupies the wavelengths from ~1530 nm to ~1610 nm, divided into a conventional band (C) 1530–1560 nm and a long band (L) 1560–1610 nm. CWDM ranges from 1271 nm to 1611 nm, and is also subdivided into a number of sub-sections, as illustrated in Fig. 10, together with the attenuation range that varies with wavelength.

An important metric of WDM is spectral efficiency SE, which is equal to the data rate divided by the spectral bandwidth SBW necessary to transmit this signal: SE (bit/s/Hz) = DR (bit/s)/SBW(Hz). Since DR = n times BR, SE becomes equal to BR divided by the channel spacing in Hz. For example, 10 Gb/s transmitted over 100 GHz bandwidth gives SE = 0.1 bit/s/Hz (very low), whereas 100 Gb/s transmitted over the same bandwidth gives SE = 1 bit/s/Hz (high). The goal is to increase SE as much as possible: a value as high as 1.32 bit/s/Hz has been achieved in a trial. The physics behind these numbers emphasizes that in reality there is a tradeoff between the BR of an individual channel and the spacing between two adjacent channels.

The WDM technique comes with its own problems, too. For one thing, we cannot launch as many channels as we want because the high power associated with many channels produces an unacceptable amount of nonlinear phenomena in the optical fiber. Another restriction on the number of channels derives from the fact that all WDM channels compete for the gain of an optical amplifier.[10]

- *Optical communications – networks*

A network differs from a point-to-point link by its ability to deliver a signal to every node out of many nodes or customers. This capability clearly indicates the need for switching and transmission protocols. A telecommunications network is the symbiosis of the physical and logical (intelligent) layers, very much similar to a computer, which is the symbiosis of hardware and software. The network's physical layer is responsible mainly for transmission and switching, whereas the logical layer is responsible for implementing the network protocols. Obviously, the better the components, the better the network, but bear in mind that even ideal components cannot change a network's topology or architecture.

There are two main network architectures today: circuit-switching and packet-routing. They are compared schematically in Fig. 11. The circuit-switching network establishes a dedicated line between end-users for the duration of their communication. This is how the public-switched telephone network (PSTN) still operates. The main drawback of this network is its very inefficient use of bandwidth. In contrast, the packet-routing network disassembles a message into separate packets, transmits these packets via any available route, and reassembles the packets into the original message before delivering it to an end-user. This transmission method is also known as IP, standing for Internet protocol, because the Internet operates by packet routing. Obviously, IP transmission maximally utilizes the network's capacity; this is why the today's trend in telecommunications is to move all types of transmission (voice, video, and data) in IP mode.

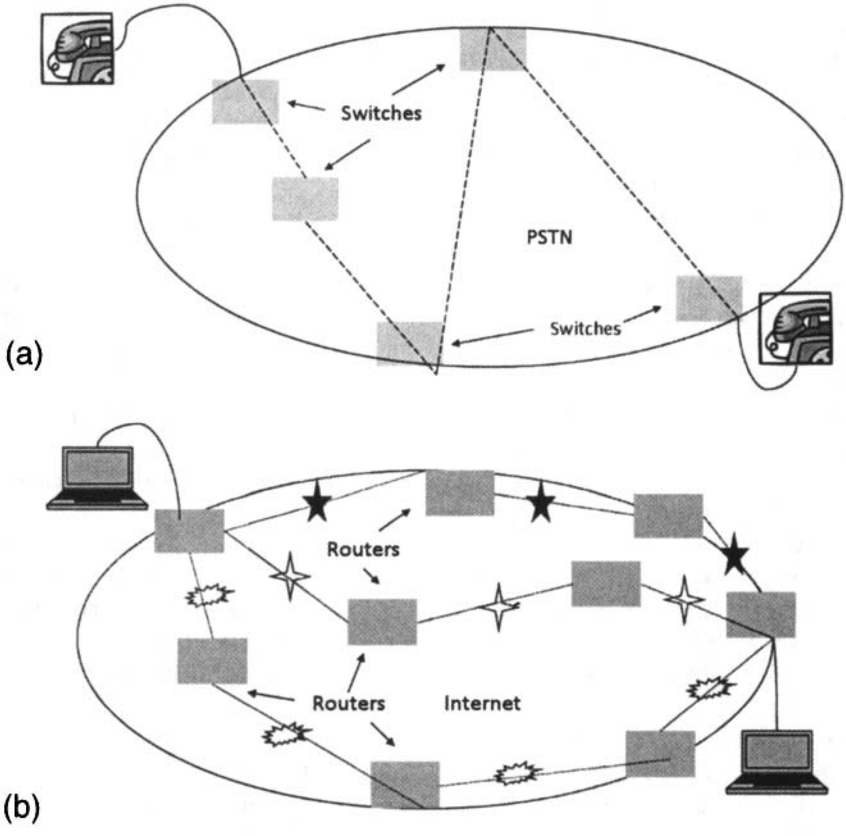


Figure 11. (a) Circuit-switching and (b) packet-routing network architectures.

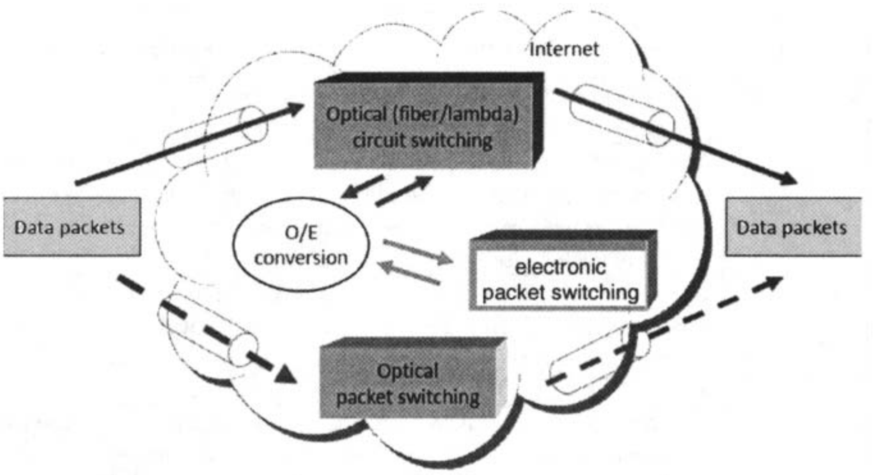
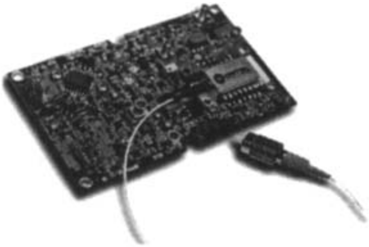
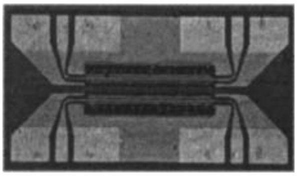


Figure 12. Internet traffic is delivered optically but routed electronically.

Optical networks are the backbone of the Internet; however, to support IP transmission over the optical network, we have to convert an optical signal into an electrical one and process (route) the packets electronically, converting them back into optical format for further transmission – see Fig. 12. This E/O/E conversion causes delays and significant expense; thus, the absence of optical packet routers is the main drawback of the optical network today. In spite of tremendous R&D efforts, optical packet routers are still far from a commercial reality – this is one of the major challenges confronting photonics researchers. There are many palliative solutions to this problem, such as using protocol-based approaches (specifically, generalized multiprotocol label switching) or employing hybrid networks in which all-through traffic is transmitted optically and add-drop messages are routed electronically. Nevertheless, IP over WDM transmission – which is how the optical-communications industry refers to it – still awaits all-optical implementation.

- *Main problems of optical communications and their possible solutions*

At this point we will consider the main problems of optical communications and their possible solutions from a top-down perspective, where *top* refers to a network and *down* to an individual component. From a physical layer standpoint, the network is a set of nodes connected by links. Nodes include systems, such as switches/routers, optical cross-connects, optical add/drop multiplexers, optical amplifiers, regenerators, *etc.* Systems, in turn, include modules, such as transmitters (Tx), receivers (Rx), dispersion-compensation modules, wavelength converters, *etc.* Modules include components, such as laser diodes, photodiodes, modulators, couplers/splitters, MUX and DEMUX, optical filters, attenuators, gratings, *etc.* The links include optical fiber and waveguides.[10] Table 2 presents some of the strategic problems and their possible solutions.

Level	Problems	Possible solutions
Network	Traffic delivered optically, but has to be routed electronically.	All-optical processing. Other?
	The cost of switching and transmission equipment, including energy bills, is currently the main barrier for the further growth of the Internet.	Energy-optimized network design,[11] lower-power-consumption equipment. Other?
	Signal distortion compensation by special systems: optical amplifiers, regenerators, dispersion-compensation devices.	New modulation schemes, new protocols, better systems and modules. Better means of transmission?
	Specific problems of PON networks: restricted distance, no switching and no protection and restoration ability (low resilience).	All-optically controlled systems, active networks based on new generation of components.
Systems and modules	Electrical-to-optical (E/O) and optical-to-electrical (O/E) conversion in routers and regenerators.	Optical packet routing, all-optical regenerators. Integrated systems?
	Switches, optical cross-connects, reconfigurable optical add/drop multiplexers, optical amplifiers, *etc.* are built from individual components with E/O/E conversions.	Miniature, integrated, all-optical modules. Integrated with transmission links?
Components	Bulky units packed from individual units. 	Monolithic integration of "optical chips". 
	High cost of manufacturing.	Silicon photonics? Easy manufacturing and easy integration with electronics, but ... bandgap problems, heat dissipation and other compatibility issues.[8]
	High cost of maintenance.	Miniature, integrated, low-power-consumption devices.

Links	Optical SiO_2 fiber: losses cause attenuation; dispersion causes decrease in bandwidth by broadening the pulse; nonlinear effects (mainly self- and cross-phase modulations) cause more dispersion and distort the waveform.	Better silica fiber. Another type of fiber?
	Waveguides for optical interconnects: the tradeoff between silicon and polymer waveguides in density and propagation delay per unit length.[12]	Improvement of existing materials? Other waveguide materials? Plasmonic waveguides?

Table 2. Problems facing optical communications and their possible solutions.

3. What will modern photonics contribute to the development of future optical communications technology?

Through research in photonics, we will discover new phenomena and new approaches to the development of devices and components that will be used in optical communications. Today's research in photonics lays the foundation for physical layer of future optical networks. But which specific research results are likely to contribute specific devices and components for tomorrow's optical communications networks is another matter. Nobody has a crystal ball and the goal of this chapter is simply to bring to the attention of the photonics research community the technology needs of the optical communications industry. Let us consider, from this standpoint, several examples of current research programs in photonics.

- *Plasmonics*

When an electromagnetic (EM) wave falls at the interface between a dielectric and a metal, polarization charges are created at the interface. The electrons in the metal will respond to this polarization, giving rise to surface plasmons – interface modes that have combined EM-wave and surface-charge properties. When light is coupled with surface plasmons, we have surface plasmon-polaritons,[13] illustrated schematically in Fig. 13(a).

We see several potential applications for surface plasmons in optical communications. For one thing, the miniaturization of conventional photonic circuits is limited by the diffraction limit, such that the minimum feature size is of the order of a wavelength. Using surface plasmons, it may be possible to overcome the diffraction limit, which can lead to the miniaturization of photonic circuits. In general, using surface plasmons can open the door to practical applications of nano-optics, where the main problem today is delivery and concentration of optical radiation energy caused by the ratio of optical wavelength (hundreds of nm) to characteristic length of nano-optical devices (nm).[13]

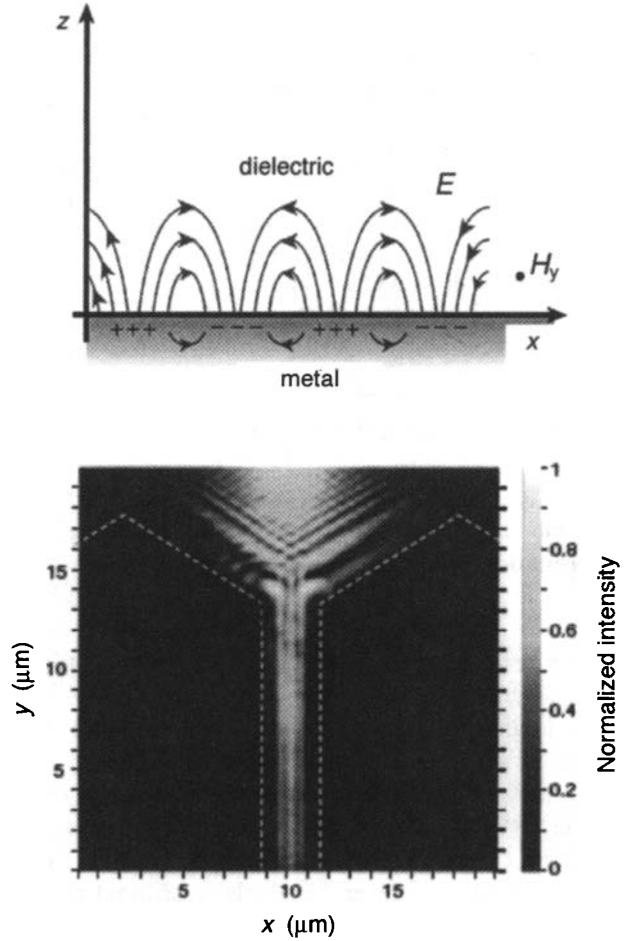


Figure 13. (a) Schematic illustration of a surface plasmon; (b) light propagation in a plasmonic waveguide.[13]

Figure 13(b) demonstrates light propagation in a plasmonic waveguide that could potentially serve in on-chip optical interconnects. A waveguide, over a centimeter in length with low bending loss has been demonstrated.[14]

Overcoming the diffraction limit with the help of surface plasmons paves the way for beam shaping, making it possible to design near-field and far-field patterns almost at will, which would allow for optical processing and coupling light into nanostructures.[15] Some recent results in surface-plasmon research of particular relevance to the optical communications community include:

- the potential of surface plasmons to be switched at THz rates and consume mJ/cm^2 power for modulation, which might lead to the creation of new types of modulators and all-optical switches;

- the fabrication of plasmonic LEDs that approach 10-GHz modulation bandwidth; this might lead to using LEDs for optical interconnects;
- the creation of a visible plasmon laser, which in turn could result in the development of new light sources for nanoscale optics;
- the development of a plasmonic lens that greatly increases photodetection capability and offers the potential for new types of photodetectors.

In general, surface plasmons engender new hope for the development of integrated devices that will perform all of their operations optically.

- *Photonic crystals (semiconductors of light)*

Photonic crystals are structures consisting of alternating thin layers of two dielectric materials with different refractive indexes; these layers reflect radiation within a resonant band and transmit other wavelengths.[16] The resonantly reflected band of wavelengths is known as the "photonic bandgap", in analogy to the electronic bandgap. The value of the photonic bandgap depends on two parameters of the dielectric layers: their thicknesses and refractive indexes. Since photons cannot propagate through solid matter if their energies fall in the photonic bandgap, photonic crystals can confine light in three dimensions.

The main applications of photonic crystals in optical communications are associated with photonic-crystal fibers. These fibers cannot serve as transmission media because they introduce huge losses, but – thanks to their special properties – they can be used for building fiber lasers, amplifiers, and nonlinear devices, which could provide dispersion control and compensation, filtering, and switching. In addition to the advantages provided by photonic-crystal fibers, optical communications could benefit from other results of the ongoing research in this area. For example, photonic-crystal lasers – which have a low-threshold current and high-speed modulation – could be employed in optical interconnects. What is more, photonic crystals potentially allow for creating tunable devices that would dramatically change the landscape of optical communications. They also could lead to the development of optical-processing devices, such as switches, photonic memory devices, and wavelength converters, the latter potentially operating at the single-photon level. Eventually, silicon photonic crystals could lead to integrated optical devices on the silicon-on-insulator (optical chip) platform, the holy grail of optical communications.[17]

- *Use of single photons in communications?*

Signals encoded in a single photon promise the absolute minimum in power consumption – and therefore cost – for generation, modulation, transmission, and detection. But can we ever really operate with a single photon for communications? Considering all the main operations occurring in sequence, we can say that the generation and detection of a single photon is a difficult but achievable task.[18] As for modulation, recently the modulation of a single photon with an arbitrary waveform was reported by Kolchin and co-workers.[18,19]

Also, an experiment confirming the feasibility of establishing a link between an orbiting source of single photons and a ground telescope was conducted recently. In this experiment, a single photon was sent and received over 150 km. Weak laser pulses were directed toward orbiting satellites that mirror optical signals on an exact parallel return path to the source. The chosen transmission parameters caused less than one photon per pulse to be returned on average, thus simulating a single-photon source. Because the signal traveled through only 8 km of the Earth's atmosphere out of a total distance of 160 km, the link has been established without serious atmospheric or geographic disturbances.[20]

In the transmission of a signal photon over a photonic crystal fiber, nonlinear effects were observed at the lowest power level, a finding that may lead to the development of all-optical processing devices.[21]

Clearly, it is impossible to even mention all the areas of photonics research under way today and obviously, it is a long way from laboratory experiments to practical technology; however, the described results provide hope and a direction that further efforts will take.

4. Conclusions

Tremendous progress in all areas of optical-communications technology has been achieved over the last two decades, including:

- significant progress in the integration of components and modules;
- advances in sophisticated and efficient modulation techniques and devices;
- transmission speeds of tens of Tb/s with spectral efficiency approaching its fundamental limit due to advances in WDM technology;
- dramatic reduction of cost per bit, which continues to decrease, making optical communications competitive at smaller and smaller distances.

At the same time, modern photonics is constantly discovering new phenomena and processes, some examples of which we briefly discussed above.

Our discussion shows that photonics and optical communications benefit from mutual fruitful cooperation. The linkup of these two fields promises progress towards the hypothetical ideal optical communications system: one that works on any scale, at any transmission speed, and over any part of the spectrum. Figure 14 illustrates this point by showing an ideal optical communications network that accepts a signal in any format and, after changing it to an optical form, seamlessly delivers the signal to any required point and in any required format. On the other hand, the research currently in progress in various photonics areas will bring new ideas and concepts to a technology that will one day boast such an ideal optical network.

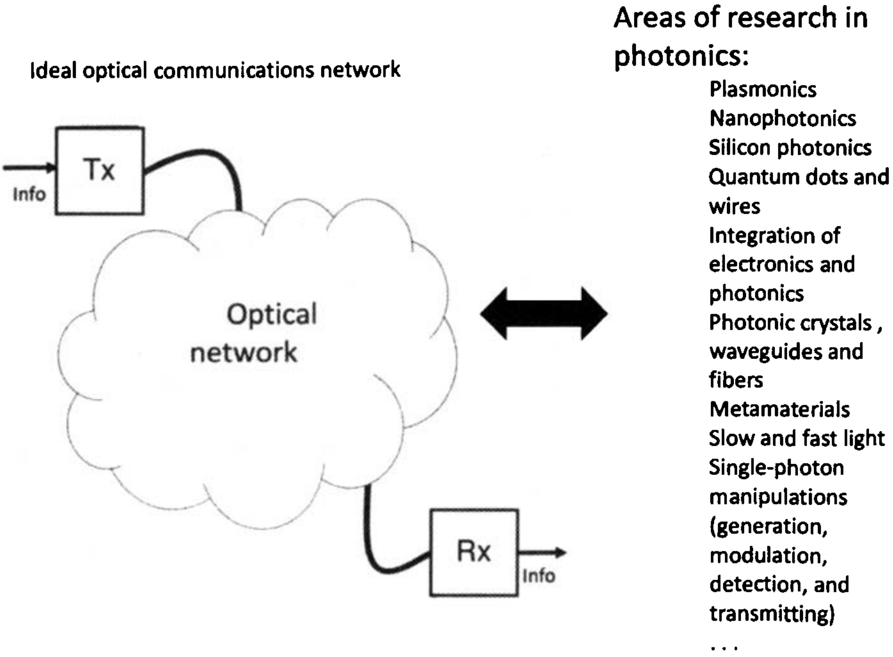

Ideal optical communications network
Tx
Info
Optical network
Rx
Info
Areas of research in photonics:
Plasmonics
Nanophotonics
Silicon photonics
Quantum dots and wires
Integration of electronics and photonics
Photonic crystals, waveguides and fibers
Metamaterials
Slow and fast light
Single-photon manipulations (generation, modulation, detection, and transmitting)
. . .

Figure 14. Ideal optical network and areas of research in photonics.

Optical communications and fundamental research in photonics are closely interconnected and therefore inspire and motivate researchers in these respective fields to pursue further advances. Only time will tell what results of today's photonics research will turn up in tomorrow's optical communication products.

References

1. Djafar K. Mynbaev and Lowell L. Scheiner, *Fiber-Optic Communications Technology*, Upper Saddle River, NJ: Prentice Hall, 2001.
2. A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, and M. B. Ritter, "Exploitation of optical interconnects in future server architectures," *IBM J. Res. Develop.* **49**, 755 (2005).
3. P. Chanclou, F. Payoux, T. Soret, *et al.*, "Demonstration of RSOA-based remote modulation at 2.5 and 5 Gb/s for WDM PON," *Proc. Optical Fiber Commun. Conf.* (2007), art. 4348854.
4. I. T. Young, E. Mohammed, J. Liao, *et al.*, "Optical I/O technology for tera-scale computing," *Tech. Dig. ISSCC* (2009), art. 4977511.
5. Y. Vlasov, "Silicon photonics for next generation computing systems," *www.ecoc2008.org/documents/SC2_Vlasov.pdf*

6. K. Takeuchi, "Technical trends in optical interconnection technology – Towards its implementation in the 'Keisoku' supercomputer," *http://www.nistep.go.jp/achiev/ftx/eng/stfc/stt020e/qr20pdf/STTqr2002.pdf*
7. G. Overton, "Can optical integration solve the computational bottleneck?" *Laser Focus World* **45**, 64 (March, 2009).
8. B. Jalali, S. Fathpour, and K. Tsia, "Green silicon photonics," *Optics Photon. News* **20**, 18 (June, 2009).
9. P. R. A. Binetti, J. Van Campenhout, X. J. M. Leijtens, *et al.*, "An optical interconnect layer on silicon," *Proc. 13th Europ. Conf. Integr. Optics* (2007), pp. 1–3.
10. Djafar K. Mynbaev, "The physical layer of the optical networks: Devices and subsystems," *Online Tutorials IEEE Commun. Soc. (2006), www.comsoc.org/tutorialsnow*
11. G. Shen and R. S. Tucker, "Energy-minimized design for IP over WDM networks," *J. Optical Commun. Networking* **1**, 176 (2009).
12. M. Haurylau, G. Chen, H. Chen, *et al.*, "On-chip optical interconnect roadmap: Challenges and critical directions," *IEEE J. Selected Topics Quantum Electronics* **12**, 1699 (2006).
13. I. Zozoulenko, "Surface plasmons and their applications in electro-optical devices," Linköping Univ. (2006), *http://www.itn.liu.se/meso-phot*
14. S. Park, "Polariton waveguide has low bending loss," *Laser Focus World* **45**, 9 (February, 2009).
15. F. Capasso, N. Yu, E. Cubukcu, and E. Smythe, "Using plasmonics to shape light beams," *Optics Photonics News* **20**, 22 (May, 2009).
16. J. Heght, "Photonic crystals foster a burst of innovation," *Laser Focus World* **45**, 103 (January, 2009).
17. C. Lopez, "Fabrication techniques for silicon photonic crystals," *Optics Photonics News* **20**, 28 (January, 2009).
18. P. Kolchin, C. Belthangady, S. Du, G. Y. Yin, and S. E. Harris, "Single photons with arbitrary waveforms," *Opt. Soc. Amer. CLEO/QELS* (2008), post-deadline paper QDPA3.
19. P. Kolchin, C. Belthangady, S. Du, G. Y. Yin, and S. E. Harris, "Electro-optic modulation of single photons," *Phys. Rev. Lett.* **101**, 103601 (2008).
20. P. Villoresi, R. Ursin, and A. Zeilinger, "Single photons from a satellite: Quantum communications in space," *spie.org/x33629.xml?ArticleID=x33629*
21. N. Matsuda, R. Shimizu, Y. Mitsumori, H. Kosaka, and K. Edamatsu, "Observation of optical-fibre Kerr nonlinearity at the single-photon level," *Nature Photonics* **3**, 95 (2009).

Ultrafast Nanophotonic Devices For Optical Interconnects

N. N. Ledentsov[*]
VI-Systems GmbH, Hardenbergstrasse 7, D-10623 Berlin, Germany

1. Introduction

The ever-growing serial transmission speed of electrical interfaces is primarily driven by an increase in the microprocessor bandwidth caused by silicon scaling. As the density of transistors doubles every two years, the processor bandwidth increases, and the speed of the input/output (I/O) ports must increase. Moore's Law for data communications predicts that a four-fold increase in the I/O speed is needed each 4–5 years. This scaling of speed imposes severe challenges to both copper and standard low-cost optical interconnects, which are presently limited to bit rates of ~10 Gb/s.

At the same time, according to the IEEE Ethernet Roadmap (Fig. 1) the single I/O bit rate should approach 100 Gb/s by 2017 while the aggregated transmission rate will increase to 5 Tb/s. At I/O speeds of 5–10 Gb/s, copper cables and connectors are becoming too energy consuming and bulky, so a massive deployment of optical interconnects is already starting.

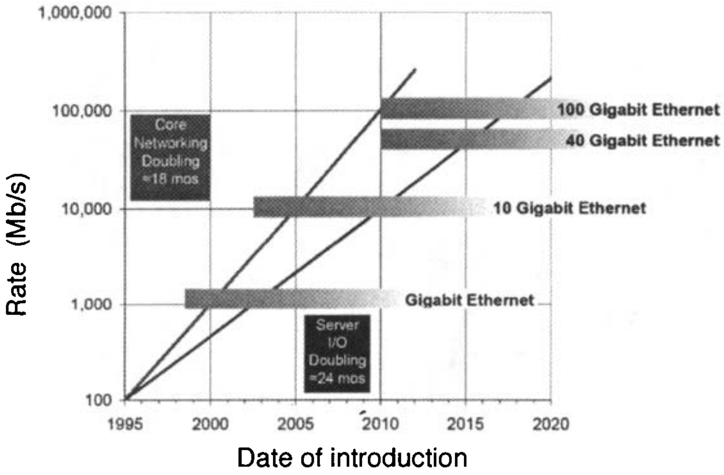


Figure 1. IEEE Roadmap for the core and I/O bit rates.

---

[*] On leave from Ioffe Physical-Technical Institute of Russian Academy of Sciences, St. Petersburg 194021, Russia.

Copyright © 2010 John Wiley & Sons, Inc.

2. Recent progress in optical components

Very short-range optical links are presently served by multimode vertical-cavity surface-emitting lasers (VCSELs) as transmitters and PIN photodetectors as detectors. Previously it was believed that going beyond 10 Gb/s with VCSELs unavoidably causes problems with reliability and spectral quality. As a result, for example, the 17 Gb/s VCSEL bit rate was abandoned in the recent 2009 fiber channel standard in favor of a reduced 14 Gb/s rate.

To overcome the problems of conventional VCSELs several approaches have been proposed:

- "antiwaveguiding" VCSEL designs that suppress in-plane microdisk modes competing for the gain and also increase the strength of the main VCSEL transition;[1]
- the use of 850 nm quantum dots (QDs) that help to increase the differential gain, while localization of nonequilibrium carriers on QDs improves device reliability;
- advanced mask and oxide aperture designs enable devices with extremely low electrical parasitics, including capacitances as low as < 90 fF.

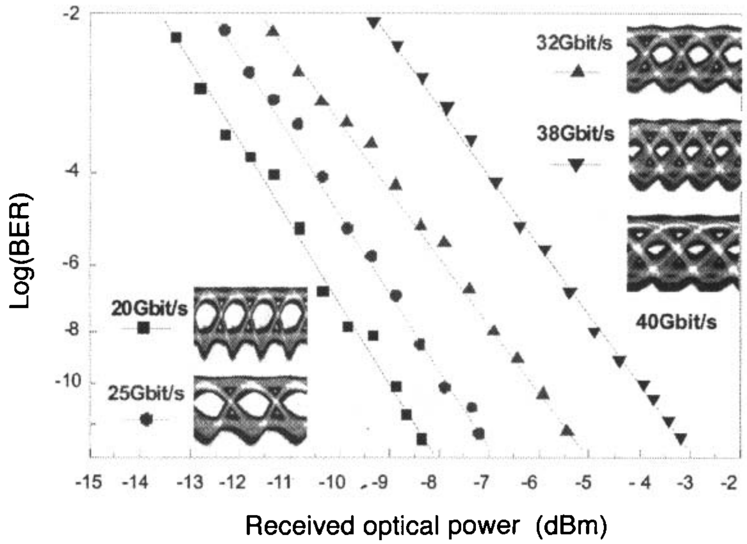


Figure 2. Bit error rate (BER) measurements for 9 μm-diameter oxide-confined VCSEL at a bias current of 9 mA and $T = 20$ °C. Insets: corresponding optical eye diagrams. Current density is ~5 kA/cm^2 and the root-mean-square (RMS) spreading of the emission spectrum is < 0.3 nm. The measurements were made with 30 GHz receiver provided by VI-Systems GmbH. "Open eye" means that "1" and "zero" signals are well separated for error-free transmission.

These new approaches have enabled VCSELs approaching 40 Gb/s bit rates at moderate current densities, shown in Fig. 2. Furthermore, these VCSELs appear sufficiently robust against degradation.[2-4] Deconvolved VCSEL rise and fall times below 10 ps have been manifested up to 100 °C heat sink temperatures.

3. Novel modulation schemes

The use of VCSELs at speeds close to or beyond 40 Gb/s may still prove challenging. A reduced spectral width is needed at ultrahigh-speed transmission due to the fiber wavelength dispersion. High-speed single-mode VCSELs are based on very small oxide-confined apertures, operate at high current densities and are not reliable. Also, strategically we should look at solutions that keep the total transceiver power consumption below 2–4 mW/Gb/s for the complete optical link.

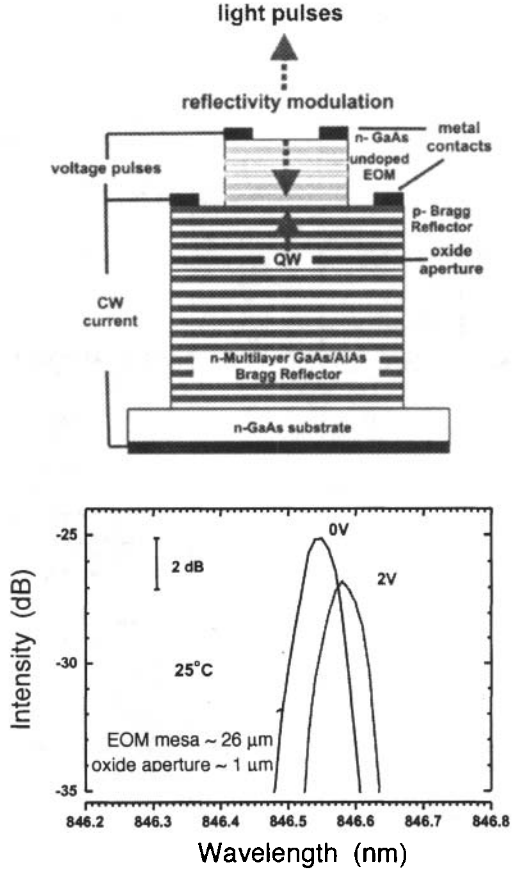


Figure 3. (a) Schematic view of an EOM DBR VCSEL device. The VCSEL section is operated in cw mode, whereas the EOM DBR section modulates the optical output signal; (b) electroluminescence intensity of a single-mode device as a function of applied reverse bias.

To fabricate VCSELs that can meet all these requirements we have proposed a vertical integration of a VCSEL with an electrooptic modulator (EOM), shown schematically in Fig. 3. The modulator section works under reverse bias, has ultralow capacitance and effectively serves as a "gate" to change the reflectivity of the top distributed Bragg reflector (DBR) section.[5,6] Recently,[6] we realized such an EOM DBR VCSEL, operating at a modulation voltage of ~2 V or less over a broad temperature range (20–100 °C). A modulator driver IC for such a device having a 100 fF capacitance may consume <1 mW per Gb/s and the development of 40 Gb/s drivers and amplifiers is in progress.

At 2 V modulation, the extinction ratio is ~2 dB for a 1 μm aperture single mode devices, see Fig. 3(b). It increases to 3–5 dB in case quasi-single-mode or multimode devices at larger apertures. At high temperatures (85–100 °C) the extinction ratio increases further, by ~30%.

The electroluminescence and optical power of the integrated EOM DBR VCSELs are shown in Fig. 4. The prototype devices are capable of open eye data transmission at ~10 Gb/s, see Fig. 5. A single-mode EOM DBR VCSELs capable of 40 Gb/s transmission at low modulation voltage is presently in development.

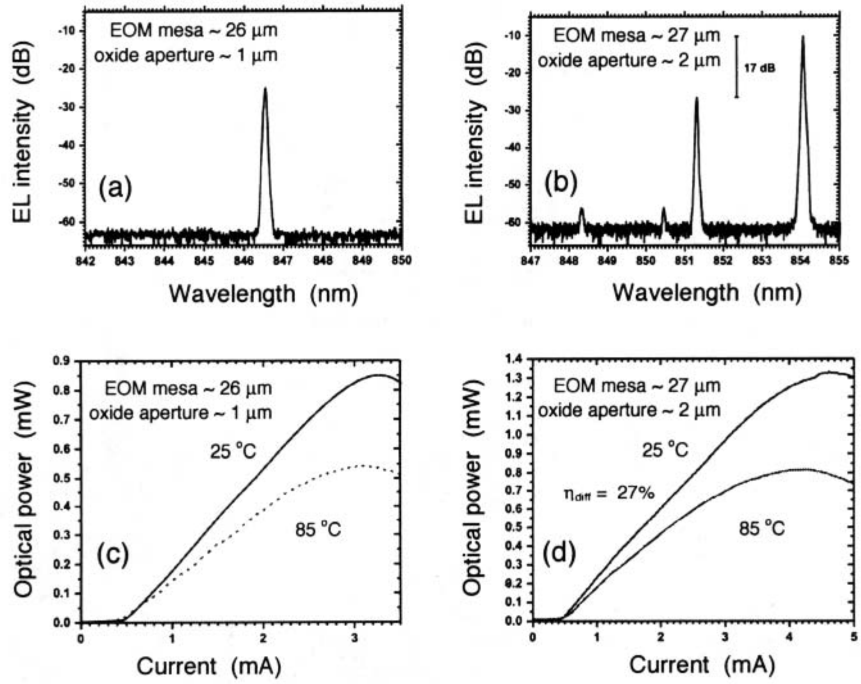


Figure 4. Electroluminescence spectra at 2 mA (a) and 3 mA (b) current drive, together with light–current characteristics of the EOM DBR devices operating in single-mode (a, c) or quasi-single-mode (b, d) regimes with 1 μm and 2 μm oxide-confined apertures, respectively.

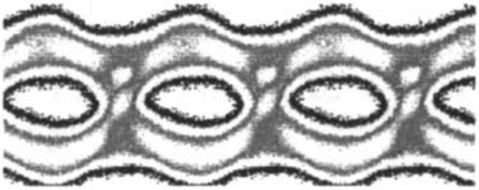


Figure 5. An open-eye diagram of the EOM DBR VCSEL with a 25 μm modulator mesa diameter (10 Gb/s, 25 °C). Average total power ~ 1 mW at ~1 V modulation voltage.

4. Integration of CMOS and III-V technologies

At bit rates exceeding 40 Gb/s the maximum signal distribution distance over a printed circuit board is limited to only a few centimeters or even less. Thus, optical transmitters and detectors must be placed as close as possible to the drivers and amplifier ICs, respectively. Preferably, the optical components must be directly integrated with the ICs.[7] As illustrated in Fig. 6, the planar design of VCSELs and top-illuminated photodetectors provides a unique vertical integration opportunity within the standard CMOS technology, because wafer-separated devices can be effectively "printed" onto CMOS ICs. Further planarization and dielectric sealing will enable their integration into a scalable, massively parallel CMOS technology.[7]

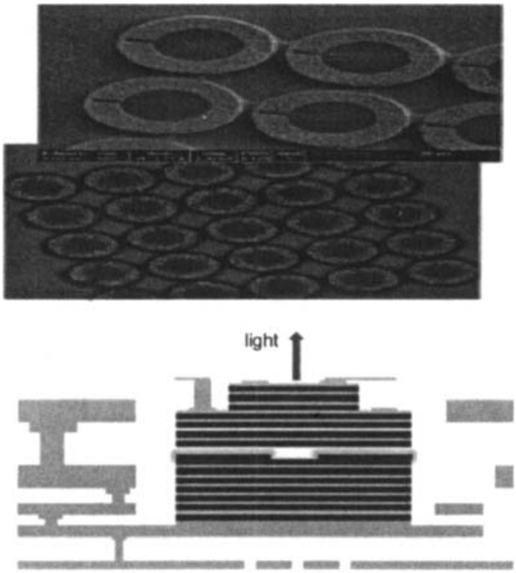


Figure 6. (a) Wafer-separated VCSEL membranes hold promise of reliable cw operation when attached to a silicon wafer; (b) the small (~5 μm) membrane thickness is compatible with CMOS-based planarization technology.

4. Conclusion

We believe that VCSELs leveraged by novel nanophotonic concepts will provide a reliable source of interconnect technologies in the near future.

Acknowledgements

This chapter was written in cooperation with J. A. Lott and V. A. Shchukin (VI-Systems GmbH); A. Mutig, G. Fiol, T. Germann, and D. Bimberg (TU–Berlin); and A. Blokhin, L. Ya. Karachinsky, M. V. Maximov, and A. M. Nadtochiy (Ioffe Physical-Technical Institute and the Academic University of St. Petersburg, Russia).

References

1. J. Humphreys and J. Scaramella, "The impact of power and cooling on data center infrastructure," IDC document #201722 (2006).
2. N. N. Ledentsov, J. A. Lott, V. A. Shchukin, *et al.*, "Quantum dot insertions in VCSELs from 840 to 1300 nm: Growth, characterization, and device performance," *Proc. SPIE* **7224**, art. 72240P (2009).
3. J. A. Lott, V. A. Shchukin, N. N. Ledentsov, *et al.*, "20 Gb/s error-free transmission with ~850 nm GaAs-based vertical cavity surface emitting lasers (VCSELs) containing InAs-GaAs submonolayer quantum dot insertions," *Proc. SPIE* **7211**, art. 721114 (2009).
4. A. Blokhin, J. A. Lott, A. Mutig, *et al.*, "850 nm VCSELs operating at bit rates up to 40 Gb/s," *Electronics Lett.* **45**, 501 (2009).
5. V. A. Shchukin, N. N. Ledentsov, J. A. Lott, *et al.*, "Ultrahigh-speed electro-optically modulated VCSELs: modeling and experimental results," *Proc. SPIE* **6889**, art. 68890H (2008).
6. N. N. Ledentsov, "Dots speed datacom VCSELs," *Compound Semicond.* **15**, 25 (2009).
7. J. A. Lott, "Fabrication and applications of lift-off vertical cavity surface emitting laser (VCSEL) disks," *Proc. SPIE* **4649**, 203 (2002).

Intersubband Quantum-Box Lasers: Progress and Potential as Uncooled Mid-Infrared Sources

D. Botez, G. Tsvid, M. D'Souza, J. C. Shin, Z. Liu, J. H. Park, J. Kirch, L. J. Mawst, M. Rathi, T. F. Kuech
Reed Center for Photonics, University of Wisconsin, Madison, WI 53706, U.S.A.

I. Vurgaftman, J. Meyer
Naval Research Laboratory, Washington, DC 20375, U.S.A.

J. Plant, G. Turner
MIT Lincoln Laboratory, Lexington, MA 02420, U.S.A.

P. Zory
University of Florida, Gainesville, FL 32611, U.S.A

1. Introduction

Semiconductor lasers operating in continuous wave (cw) at or near room temperature (RT) and emitting in the mid- and far-infrared wavelength ranges, 3–13 μm, are critically needed for a vast array of applications. Intersubband (IS) transition emitters are the most likely solution. The first implementation of the concept for using IS transitions for laser action was realized in early 1994[1] and named quantum cascade (QC) laser. Current QC devices have demonstrated high-power, RT cw operation in the 4.0–6.0 μm range,[2,3] but with relatively low wall-plug efficiency (~10%) due to inherently high operating voltages (10–11 V). Furthermore, the devices have highly temperature-sensitive characteristics[2-4] at and above RT, due to severe carrier leakage out of their active regions, which in turn raises serious issues of long-term device reliability. To suppress the carrier leakage we have employed the deep-well QC laser concept.[5,6]

Quantum cascade lasers have fundamentally poor radiative efficiencies since the LO-phonon-assisted relaxation time for electrons in the upper lasing states is about 2 ps, whereas the radiative relaxation time is about 4 ns. That is, nonradiative processes are about 2000 times faster than radiative processes. Since there are good reasons to believe that the LO-phonon-assisted relaxation time will substantially increase if the relaxing electrons are confined in quantum boxes,[7-15] the radiative efficiency problem can be overcome by replacing the quantum-well (QW) active regions of a QC laser with a 2D array of quantum boxes (QBs)[16,17] or of cascaded QBs.[18]

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

2. Suppression of carrier leakage in QC lasers: The deep-well concept

The active medium of conventional QC lasers is composed of QWs and barriers of fixed composition. As a result, state-of-the-art devices optimized for high cw power and emitting in the 4.5–5.0 μm range experience substantial carrier leakage from the upper laser level to the continuum. This fact is evidenced by a rapid decrease in the slope efficiency η_S above 300 K (*i.e.* the characteristic temperature T_1 for η_S has a low value of ~140 K)[3,4] and also by a relatively low characteristic temperature for the threshold-current density J_{TH} above 300 K (T_0 ~ 140 K).[2,3] The low values of T_0 and T_1 are both indirectly attributable to the rather small energy differential (~250 meV) between the upper lasing level and the top of the exit barrier.[2] Figure 1(a) illustrates the conduction-band (CB) profile, energy levels and square modulus of the relevant wavefunctions in the active region and in the upper and lower Γ minibands of the relaxation/injection (R/I) regions for a conventional λ = 4.8 μm $In_{0.68}Ga_{0.32}As/Al_{0.64}In_{0.36}As$ QC laser[2] as calculated by an 8-band ***k·p*** code. Here the active-region design is of the double phonon-resonance type and

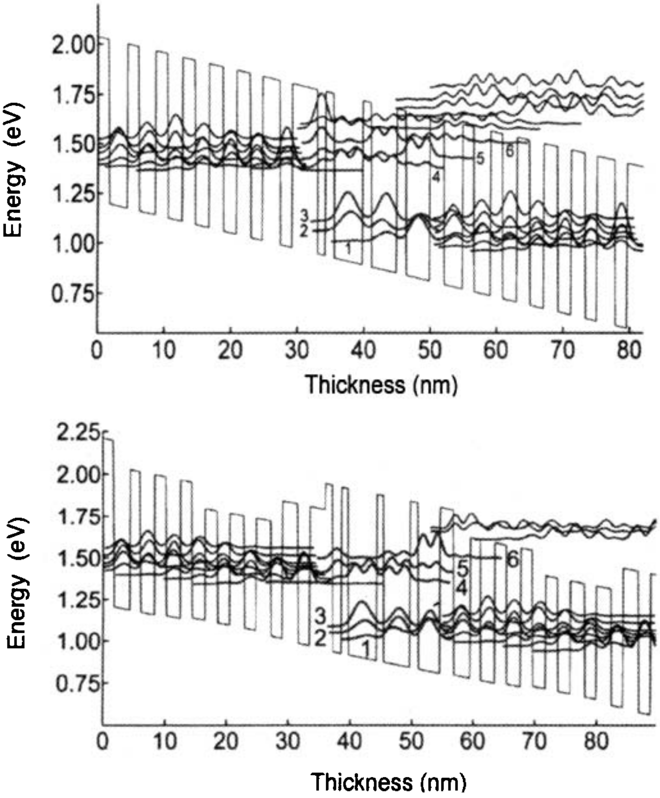


Figure 1. Conduction band diagram and relevant wavefunctions: (a) conventional QC laser (λ ≈ 4.8 μm);[2] (b) deep-well QC laser (λ ≈ 4.8 μm).[6] The upper lasing state is 4; while 5 and 6 are higher-lying states in the active region.

the lasing transition occurs between levels 4 and 3. At and above room temperature, some fraction of the carriers injected into the upper laser level (4) are thermally excited to higher active-region levels (*i.e.* levels 5 and 6 in this case). The level-6 wavefunction overlaps quite well with that of the lowest state in the R/I region's upper Γ miniband, and in turn electrons can escape rather easily to the continuum. With increasing thermal activation at high heatsink temperatures, this carrier-loss mechanism induces a rapid increase of J_{TH} and a rapid decrease of η_S, leading to relatively low T_0 and T_1 values. In contrast, the designed deep-well devices[6] shown in Fig. 1(b) use the injector regions of 5.4 μm-emitting QC lasers[19] (*i.e.* $Al_{0.56}In_{0.44}As/In_{0.60}Ga_{0.40}As$ superlattices), while the QWs in the active regions are $In_{0.68}Ga_{0.32}As$, thus deeper in energy than the QWs in the injector. Further, for strain compensation, the barrier layers in and around the active region are $Al_{0.75}In_{0.25}As$, thus taller than the barriers in conventional 4.8 μm QC lasers of Fig. 1(a). As seen from the CB diagram, the active-region upper levels (5 and 6) have wavefunctions tightly confined to the active region, due to fact that the energy difference between the upper lasing level and the exit-barrier top has increased from ~250 meV to ~450 meV, such that negligible escape to/relaxation to the upper/lower Γ miniband occurs.[5,6]

Ridge-guide devices were mounted epi-side up on Cu blocks and measured in pulsed mode (100 ns, 2 kHz). Results from 3mm-long, uncoated chips are shown in Fig. 2. The J_{TH} value of 1.78 kA/cm^2 at 20 °C is comparable to those derived for uncoated 3-mm long conventional 4.8 μm QC devices[2] with similar injector-doping sheet density (10^{11} cm^{-2}) and same number of periods. The characteristic temperature for J_{TH} – defined by $J_{TH}(T_{IN}+\Delta T) = J_{TH}(T_{IN})\exp(\Delta T/T_0)$ where $T_{IN}+\Delta T$

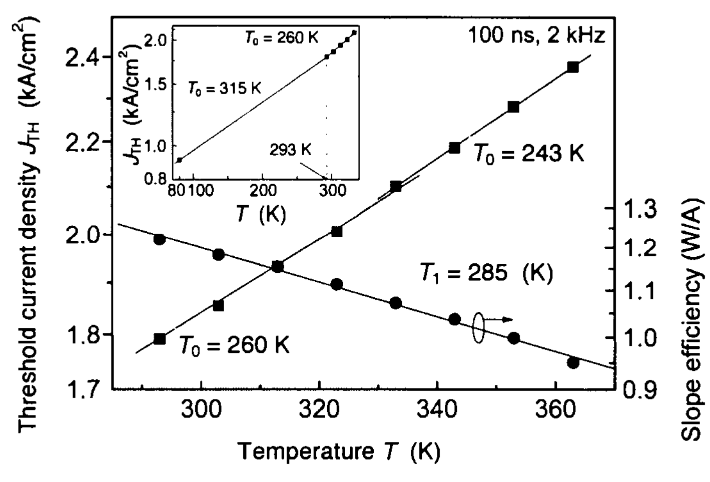


Figure 2. Threshold-current density J_{TH} and the slope efficiency *vs.* heatsink temperature *T*. The characteristics temperature coefficients T_0 and T_1 are for J_{TH} and the slope efficiency, respectively, for deep-well QC lasers($\lambda \approx$ 4.8 μm).[6] Inset: J_{TH} *vs.* *T* from 80 K to 333 K.

is the heatsink temperature T and T_{IN} is the initial T – is 260 K for 20–60 °C and 243 K for 60–90 °C. (The inset of Fig. 2 shows that $T_0 \approx 315$ K for 80–293 K.) By comparison, T_0 is only 143 K for high-performance conventional 4.6 μm QC lasers[2] over the same temperature range. Similarly, the characteristic temperature for η_S, defined by $\eta_S(T_{IN}+\Delta T) = \eta_S(T_{IN})\exp(\Delta T/T_1)$, is 285 K over 20–90 °C as compared to only ~140 K for 4.6–4.8 μm QC lasers operating over the 20-60 °C[3] and 0-50 °C[4] temperature ranges. Moreover, T_0 values as high as 278 K were observed over the 20–90 °C temperature range.[6] These ultra-high values for T_0 and T_1 mean that we are approaching temperature dependences determined mainly by optical phonon scattering. They further indicate that that leakage to the indirect valleys (X or L) is not an issue in the 4.5–5.0 μm range. The virtual doubling of T_0 and T_1 above RT should lead to significantly improved cw performance[5,6] (*e.g.* an increase in RT wall-plug efficiency from ~10 % to ~20 %), as well as to long-term reliability at watt-range cw powers. Furthermore, tight carrier confinement makes possible the realization of intersubband QB (IQB) lasers,[16-18] since poor carrier confinement would prevent long relaxation times from occurring.

3. IQB lasers

Creating QBs causes discrete states in the subbands,[16,17] which in turn causes the LO-phonon-assisted electron relaxation time to increase[7,9] by a factor β.[17] This is due to the so-called *phonon bottleneck* effect. Experimental results[10,12,13] from QBs and photocurrent-response/dark-current measurements from QB IR detectors[11,14] indicate electron-relaxation times of the order of 100 ps, in good agreement with theory.[15] In self-assembled QBs, as the temperature increases to RT, the relaxation times decrease[12] due to the inherent carrier losses to wetting layers.[20] The proposed QBs are not self-assembled and thus can be made *deep* (*i.e.* negligible carrier leakage with increasing temperature). Thus, for deep InGaAs/AlInAs/InP QBs β may well be as high as 50 at RT.

The proposed device is schematically shown in Fig. 3. An IS-transition laser with an active region composed of a 2D array of QB ministacks, called "active boxes", separated by current-blocking material, as shown in Fig. 3(a). Each ministack is composed of 2 to 4 QBs, sequentially arranged like the stages in a QC-laser structure, see Fig. 3(b). The ministacks of QBs are fabricated by dry etching and crystalline regrowth, allowing for tight carrier confinement. The active-box array, together with low-doped *n*-type InP layers, constitutes the core of an optical waveguide with heavily doped *n*-type InP cladding layers. The designed active region has deep wells in energy, which, due to the 450 meV-tall barrier for electrons in the upper energy state E_4, ensure suppression of carrier leakage.

Consider a 2D array of 4-QB stacks integrated into a designed transverse waveguide of low loss coefficient, $\alpha_W \sim 1.5$ cm^{-1}. Then, for 0.75 mm-long, 20 μm-aperture devices with 10%/90% front/back-facet reflectivities and taking $\beta = 50$, the calculated cw RT wallplug efficiency η_P shown in Fig. 4 reaches a maximum of 50% at ~15 times threshold and at a wavelength of 4.8 μm. The J_{TH} value is ~0.1

kA/cm^2 while the voltage is ~1.7 V, in sharp contrast to typical high QC-device voltages of 10–12 V. The proposed QBs do not involve self-assembly, but actual fabrication,[21] thus allowing to tightly confine the carriers to the QBs. Due to the deep-QW proposed design carrier leakage will be suppressed, as already experimentally demonstrated for deep-well QC devices.[5,6] Thus, truly deep QBs would be realized for the first time, which in turn will allow for temperature-insensitive characteristics and subsequent high wallplug efficiencies and long-term device reliability.

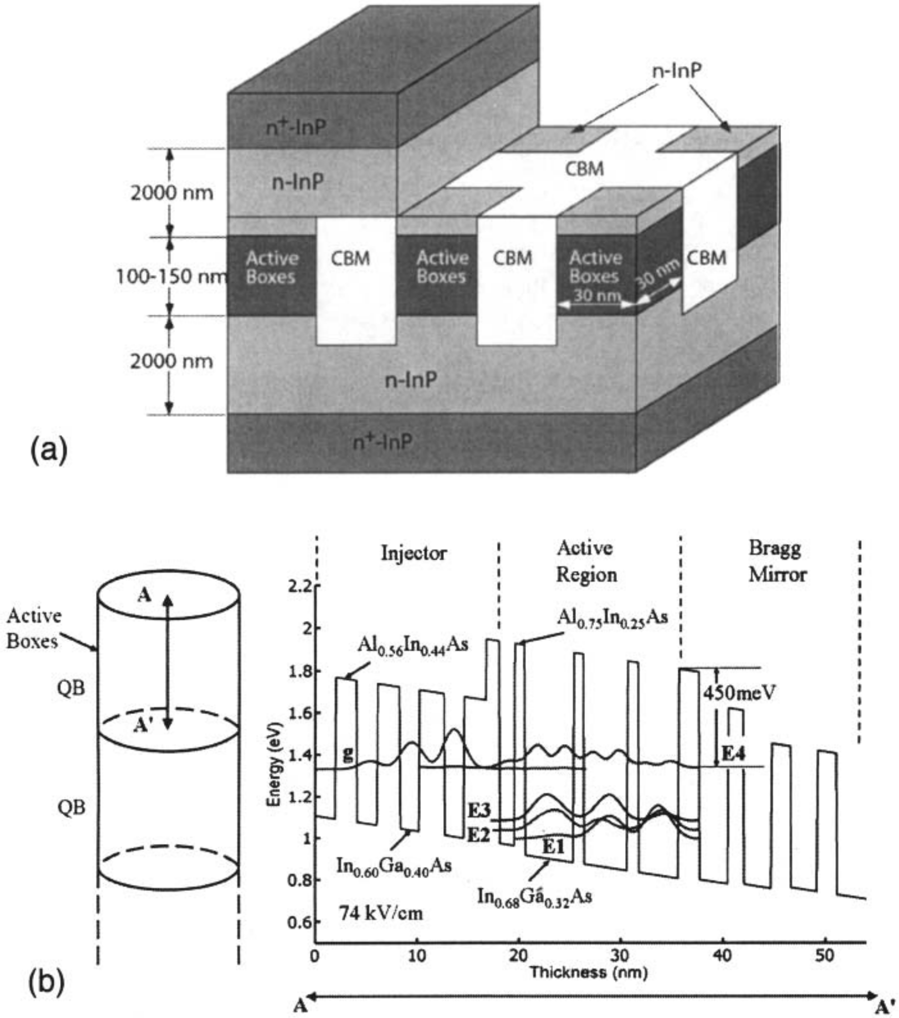


Figure 3. (a) Representation of the IQB structure (where CBM refers to current blocking material); (b) schematic representation of a QB ministack (*i.e.* active boxes) and the conduction-band energy diagram inside one QB (λ ≈ 4.8 μm), indicated by the AA' line.

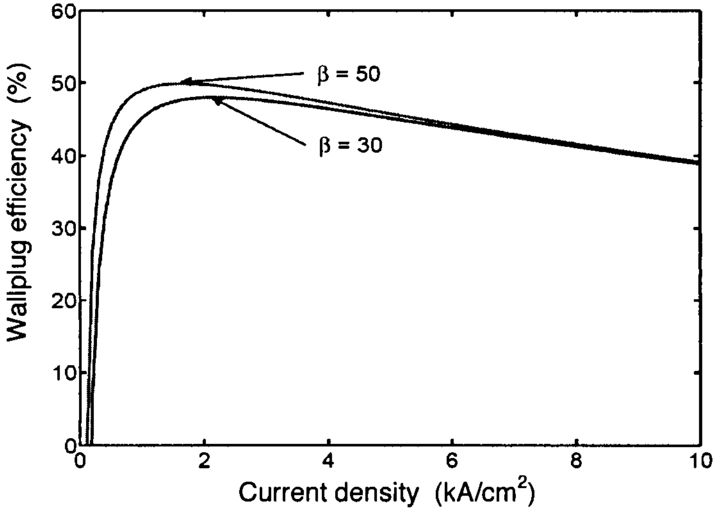


Figure 4. Calculated RT cw wallplug efficiency *vs.* current density for 4-QB-stacks devices at different β values and for $\lambda \approx 4.8$ μm.

4. Active-photonic-crystal (APC) structures: Scaling to 1 W coherent power

We have already shown that 50% wallplug efficiency is theoretically achievable by using 4QB-stack devices. However, even at an η_P value of 50%, the estimated power for a ridge-type device of 20 μm-wide emitting aperture is rather low: ~54 mW. Therefore, in order to achieve 1W output at high η_P values, we need to scale the coherent power. We propose to do this at the chip level by using an APC-type structure.

Active-photonic-crystal laser structures for watt-range coherent power were first proposed and demonstrated as early as 1989.[22] Called resonant-optical-waveguide (ROW) arrays, such devices have shown operation in near diffraction-limited beams to 10 W peak-pulsed power[23] and 1.6 W cw power.[24] In contrast to the vast majority of APC-type devices, for ROW arrays the gain is preferentially enhanced on the low-index crystal sites,[22] see Fig. 5(a). This favors lasing of only the APC modes peaked at those APC sites (*e.g.* the in-phase mode shown in Fig. 5(a)). The index step of the APC structure is typically in the 0.05-0.10 range.

To realize an APC structure of the type shown in Fig. 5(a), high-index material (such as n^--InP) is preferentially regrown to create the high-index APC sites,[25] as shown in Fig. 5(b). The regrown high-index regions "draw" the field away from the active region, thereby achieving two desired effects at once: high-effective-index regions and low modal gain due both to significantly reduced (transverse) field overlap with the active region and to strong absorption losses to the metal contact. (The transverse waveguide in the high-index regions is of the surface-plasmon type). Preferential lasing of the desired in-phase mode, shown in Fig. 5(a), occurs at and near its lateral resonance.[22] For λ = 4.8 μm, the lateral resonance is calculated to occur for s = 6.5 μm, where s is the width of the high-index region in Fig. 5(a). At and near its resonance, the in-phase mode has ~99% of its field in those regions, while the corresponding value for other modes is as low as 86%.[25] In turn, only 1% of the mode "sees" losses in the high-index regions, while the high-order modes "see" significantly more loss. Thus, lasing only in the in-phase mode is assured.

Since the device area is large (~0.3 mm^2), e-beam lithography is impractical for QB patterning. Instead we plan to use block copolymer lithography. Then the formation of 20 nm disks on ~40 nm centers becomes possible,[26] significantly increasing the probability of obtaining β values of 50.

At resonance, to get a significant amount of energy (*i.e.* ~66%) in the central far-field lobe, the array element width d should be about 5 times larger than s, see Fig. 5(a).[27] We choose $d = 35$ μm. Then, given that 54 mW are obtained from a 20 μm-wide aperture, the power per element is 95 mW. Subsequently, for a 12-element-array the power scales to 1.14 W. For computing η_P we have to take into account that the waveguide loss α_W increases from 1.5 cm^{-1} to 3 cm^{-1}. Then, as shown in Fig. 6, a maximum cw η_P value of 46 % is estimated for at an output

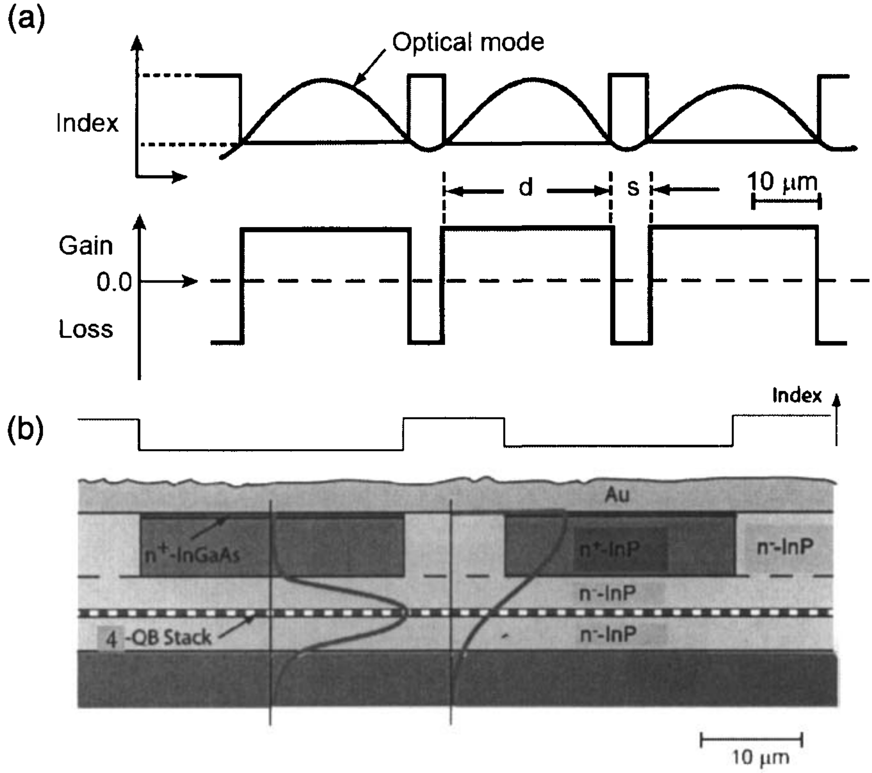


Figure 5. (a) Schematic representation of an active photonic crystal (APC) laser structure; (b) cross-section of InP-based APC structure with 4-QB-stack active region.

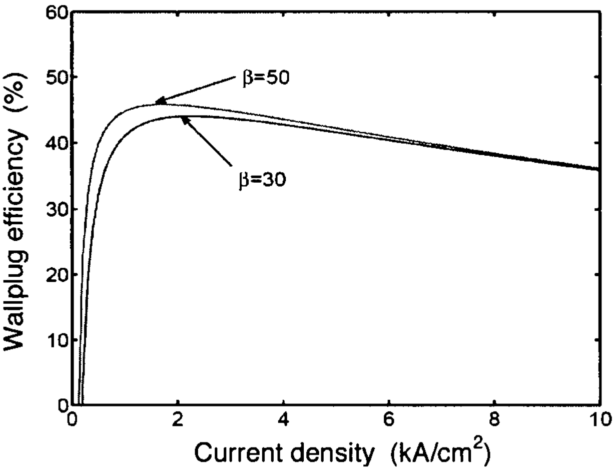


Figure 6. Calculated RT cw wallplug efficiency *vs.* current density for 4-QB-stack,12-element APC devices at different β values and for $\lambda \approx 4.8$ μm.

power of 1.16 W. Of course, since only ~66% of the power is emitted in the far-field central lobe, the useful emitted power drops to 0.77 W, thus giving an effective η_P value of 30 %. However, as long as the emitted power is spatially coherent and of flat phase front, one can use the method of amplitude-to-phase conversion^{28} to garner all light in the central lobe. Quite practical assumptions are that 95% of the power will be garnered in the central lobe and that the necessary optics will cause a 10% drop in output. Then the useful cw power becomes 1 W, delivered with a wallplug efficiency of 40%.

5. The potential of IQB lasers

The emitted power, P_{OPT} at η_{Pmax} is 54 mW. Since the heat-dissipation power P_H is simply given by $P_H = P_{OPT}[(1/\eta_P) - 1]$, P_H is only 54 mW. Figure 7 shows what happens when the number of QBs in a ministack varies from 1 to 4. The η_{Pmax} value decreases to 33%, reducing P_{OPT} to 15 mW, while the dissipated heat P_H falls to 30 mW. If β is taken to be 30 (not shown), η_{Pmax} goes down to 29%, while P_{OPT} and P_H increase to 19 mW and 47 mW, respectively. That is, for both the most ambitious case: 4-QB stacks with β = 50, and the least ambitious case: 1-QB stacks with β = 30, $P_H \sim 50$ mW, two orders of magnitude smaller that typical P_H values of 5–6 W of QC lasers emitting 20–50 mW. Such small values for P_H would definitely allow *uncooled* device operation. The only penalty for going from 4-QB-stack devices to 1-QB-stack devices is a ~3-fold decrease in optical power. For most light-absorption spectroscopy 20 mW power is adequate; thus, 1-QB-stack devices will suffice. However, when high (*i.e.* ~0.8 W) cw coherent power is needed to be delivered under uncooled conditions (*i.e.* $P_H \sim 1$ W) 4-QB-stack devices, as elements of APC structures of Fig. 5(b), will definitely be necessary.

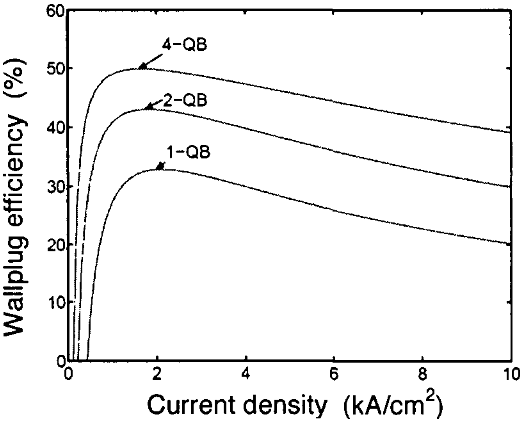


Figure 7. Calculated RT cw wallplug efficiency *vs.* current density for devices with different number of QBs in the QB ministacks, when $\beta = 50$.

The realization of uncooled 4.0–5.0 μm highly efficient cw devices would revolutionize noninvasive medical-diagnostics techniques based on laser-absorption spectroscopy, such as body-fluid and breath analysis. For breath analysis, the sensitivity would increase by four orders of magnitude to detect certain vital-organs malfunctions, metabolic disorders and (invisible) traumas; and to monitor health and fitness. In turn, doctors could use small, portable units for immediate diagnostics in their offices or during critical times such as surgery

Uncooled cw mid-IR compact sources operating at room temperature would also revolutionize the field of laser-based gas detection, because gas-sensing systems would become robust and thus practical. For example, toxic chemical agents such as sarin and s-mustard, as well as explosives such as TNT and ammonium nitrate, could be detected with lightweight, portable sensors. The other major impact would be as high-power (0.6–0.8 W) portable sources for remote sensing (*e.g.* laser photo-acoustic spectroscopy) or for IR countermeasures (*i.e.* missile-avoidance systems) for which the use of lightweight, high-efficiency, small sources is crucial.

6. Fabrication of IQB devices and preliminary light-emission results

- *Basic fabrication sequence*

The basic fabrication for the realization of a 2D QB array is schematically shown in Fig. 8. After initial MOVPE growth, the wafer is covered with a thin ~40 nm dielectric film (SiO_2 or Si_3N_4) and, after e-beam resist is deposited, the QB-pattern is defined via e-beam lithography. Then RIE is used to transfer the pattern into the dielectric layer (unlike what is shown in Fig. 8, the mask is made of disks rather than squares). The 2D array of dielectric disks is used as a mask for QB-stack formation. Then, dry etching is used to remove material followed by preferential regrowth of semi-insulating InP in an MOVPE system. After the regrowth the dielectric disks are chemically removed and regrowth is performed via MOVPE to complete the structure shown in Fig. 3(a).

- *Interface passivation*

For unipolar intersubband devices such as QC lasers one need not be concerned about loss of carriers to defects at exposed surfaces, since the transition energies involved are much smaller than the energy between midgap and the conduction-band edge at those exposed surfaces. QC lasers with exposed-surface, 10 μm-wide ridges operate quite well. However, for nanostructures with in-plane dimensions of ≤ 50 nm, the defect density needs to be drastically reduced at the device edges, since Fermi-level pinning would cause full depletion across the devices. Our QB fabrication was therefore done such that virtually no charge-trapping states are formed at the QB sidewalls, thus avoiding Fermi-level pinning. We have carried out experiments on (110)-oriented InGaAs surfaces,[29] that is, on crystalline planes equivalent to the side edge(s) of the QBs to be formed. Initially, a thick *n*-type InGaAs was grown on top of an *n*-type (110)-oriented InP wafer. Subsequently it was exposed to RIE etching in either a $CH_4/H_2/Ar/Cl_2/BCl_3$ or a $Cl_2/SiCl_4/Ar$ gas mixture that have been used to fabricate nanopoles for the IQB structures. After an annealing step and surface treatment with tartaric acid, regrowth of *n*-type InGaAs was carried out in an MOVPE reactor. Then, electrochemical capacitance-voltage (ECV) measurements were performed to check the doping profile and to determine the density of charges trapped at the interface due to surface states. As evident from Fig. 9, the apparent carrier concentration *vs.* depth curves show no abrupt features at and around the interface, which is clear indication that the interfacial defect density N_{INT} is below $10^{11}/\mathrm{cm}^2$. In turn, that means that Fermi-level pinning has been eliminated at the interface.

- *Nanopatterning and etching of nanopoles*

The eventual design involves a QB array of 30 nm diameter boxes on 60 nm centers, a nontrivial task. Therefore, the patterning has to be done in progressive steps. We used e-beam direct writing employing the hydrogen silsesquioxane resist, which was found best for generating high-quality 2D dot patterns: 33 nm diameter dots on 80 nm centers. Then we transferred such patterns into Si_3N_4, the dielectric found suitable as a mask for dry etching and regrowth.

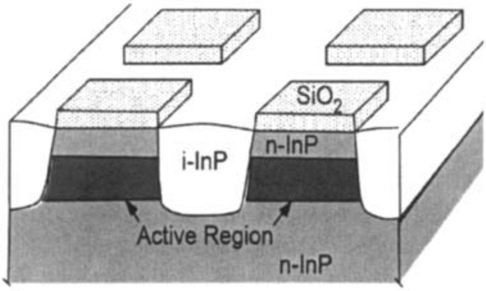


Figure 8. Schematic representation of the fabrication for the quantum-box array.

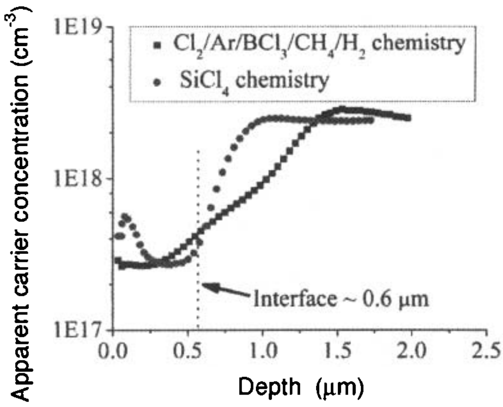


Figure 9. ECV data for $CH_4/H_2/Ar/Cl_2/BCl_3$ and $Cl_2/SiCl_4/Ar$ plasma etching.[29]

Using 40 nm thick Si_3N_4 disks, plasma etching was performed with both electron-cyclotron resonance (ECR) etching using CH_4/H_2 + Argon + BCl_3 as well as inductive coupled plasma (ICP) etching using $SiCl_4$. Using both methods we were able to obtain 30–40 nm diameter nanopoles of ~70 nm height on spacings as low as 100 nm. This corresponds to a 1-QB-stack device. A typical result from an ICP-etched sample is shown in Fig. 10(a), where the leftover Si_3N_4 regions appear as "white" caps atop the nanopoles.

- *Regrowth for electrical insulation between nanopoles*

Using dry-etched nanopoles, like the ones shown in Fig. 10(a), we performed regrowths of semi-insulating InP to ensure electrical isolation between the QBs in the nanopoles. The regrowths consisted of a thin layer of *n*-InGaAs followed by high-resistivity Fe-doped InP. A typical result of successful regrowth is shown in Fig. 10(b).

From a close inspection in the SEM, we find that the top diameters of the nanopoles are in the 27–30 nm range, which in turn provides an ~40 nm diameter at the active region. Such results have been obtained for nanopoles on 100 nm, 125 nm and 150 nm centers. The leftover Si_3N_4 caps have subsequently been removed with HF, followed by the rest of the regrowth for the formation of IQB-laser structures, as shown in Fig. 3(a).

- *Electroluminescence data*

Ridge-guide devices of 15 μm width and 4 mm length were fabricated from both IQB material and single-stage material, to be used for comparison. Single-stage devices provided electroluminescence (EL) at the designed 4.8 μm wavelength. On the other hand, the emission spectra of IQB devices contained two peaks, at 3.5 and 3.8 μm, see Fig. 11. These results represent the first demonstration of electroluminescence from QB structures fabricated by etch and regrowth. They prove that our passivation process[29] has prevented Fermi-level pinning at the nanopole sidewalls and subsequent depletion of the QBs.

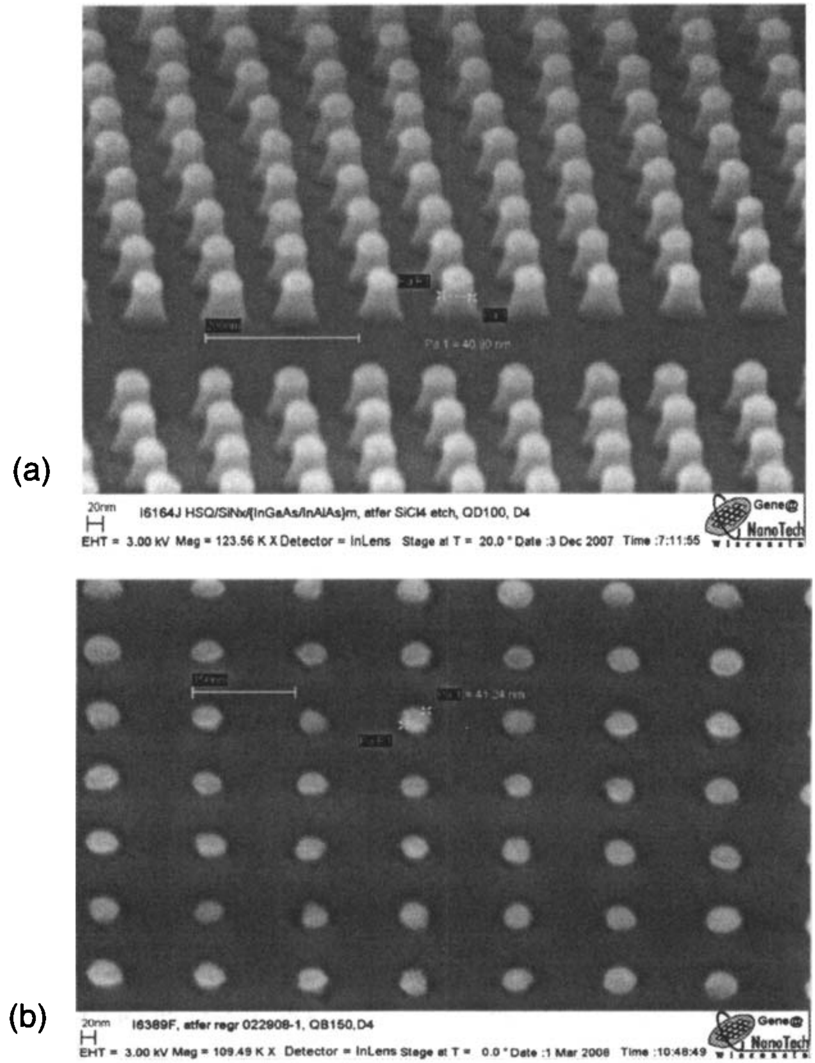

(a)
Pa 1 = 40.90 nm
20nm
I6164J HSQ/SiNx/[InGaAs/InAlAs]m, after SiCl4 etch, OD100, D4
EHT = 3.00 kV Mag = 123.56 K X Detector = InLens Stage at T = 20.0 ° Date :3 Dec 2007 Time :7:11:55
Gene@ NanoTech
(b)
= 41.34 nm
20nm
I6389F, after regr 022908-1, OB150,D4
EHT = 3.00 kV Mag = 109.49 K X Detector = InLens Stage at T = 0.0 ° Date :1 Mar 2008 Time :10:48:49
Gene@ NanoTech

Figure 10. SEM images of: (a) 40 nm diameter nanopoles on 100 nm centers dry-etched using ICP with $SiCl_4$ and using a 2D array of Si_3N_4 disks as the etch mask; (b) top view of nanopoles embedded in regrowth.

The two EL peaks correspond to transitions from levels 5 to 2 and 5 to 3, respectively, within the devices active regions. This is expected since the current densities through the QBs are much higher than those through the single-stage devices. One notable fact is that the EL spectra widths at 3.8 and 3.5 μm were comparable to the single-stage EL-spectrum width, which in turn confirms that in intersubband QBs made by etch and regrowth the emission spectra are not sensitive to variations in the QBs lateral dimensions, as expected from theory.[17]

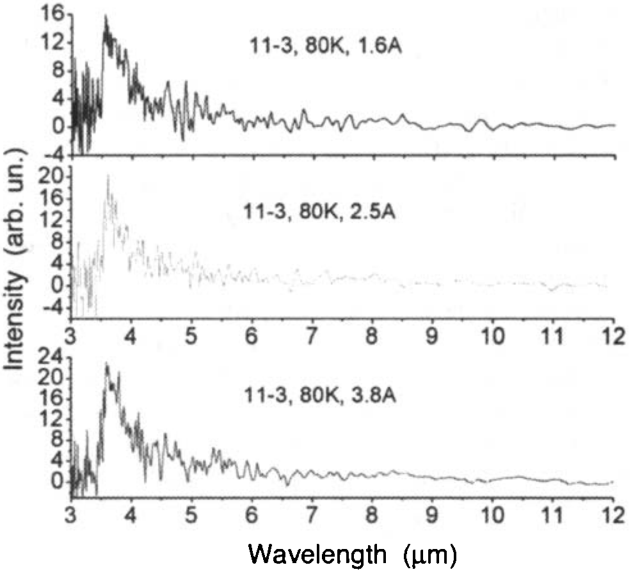


Figure 11. Electroluminescence spectra from IQB structures of 40 nm diameter on 125 nm centers at various drive levels.

However, these preliminary results were obtained from devices with abnormally high voltages: 12.5 and 14 V for single-stage and QB devices, respectively, at 1 A drive. The reasons were found to be: a) inadvertent growth of highly resistive, InP waveguide layers; and b) an oxide/film formed between the top QB InGaAs layer and the Si_3N_4 mask layer, which, after mask removal, led to high defect densities at the interface between the top QB layer and the regrown material (the latter was confirmed from ECV measurements like the ones shown in Fig. 9). Abnormally high voltages cause serious problems: impact ionization gives rise to both hot electrons, which in turn leads to carrier loss, and holes, which in turn cause bandgap emission. These significant carrier losses explain the relatively weak EL signals from both single-stage and QB devices. The causes of high operating voltage can be eliminated by: a) intentional (low) doping of the InP waveguide layers; b) the use of a sacrificial InGaAs layer. We have already proven the latter by growing two thin layers: 3 nm InP + 4 nm InGaAs, atop the QC structure to be used for IQBs, and using tartaric acid to preferentially etch the InGaAs layer. Then we found, from ECV measurements, that the trapped charges disappear.

7. Conclusions

Significant suppression of phonon-assisted relaxation times in *deep* QBs will allow the fabrication of intersubband lasers emitting cw at room temperature in the 4–6 μm wavelength range with high (45–50%) wallplug efficiency. Due to much lower heating in IQB devices compared to QC devices, it should be possible to coherently scale the power laterally via active-photonic-crystal structures. Then, up to 1 W cw power emitted in a single, diffraction-limited lobe is expected with 40% wallplug efficiency. In turn, uncooled mid-IR emitting lasers become a possibility, and then one will be able to make hand-held and/or portable sensors for mid-IR, low- *and* high-power applications. Our preliminary results include the suppression of carrier leakage in 4.8 μm QC lasers which in turn led to thresholds and slope efficiencies that vary with temperature 2.5 times *slower* than in conventional devices, the development of a passivation technique that eliminates trapped-charge defects at interfaces corresponding to the QB sidewalls, dry etching of 30–40 nm diameter nanopoles on center-to-center spacings as small as 100 nm, and successful regrowth of semi-insulating material around the nanopoles. Finally, electroluminescence was observed which represents the first time that electroluminescence is obtained from QB structures fabricated by etch and regrowth.

Acknowledgments

The authors gratefully acknowledge valuable discussions with Claire Gmachl.

References

1. J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser," *Science* **264**, 553 (1994).
2. Y. Bai, S. R. Darvish, S. Slivken, W. Zhang, A. Evans, J. Nguyen, and M. Razeghi, "Room temperature continuous wave operation of quantum cascade lasers with watt-level optical power," *Appl. Phys. Lett.* **92**, 101105 (2008)
3. A. Lyakh, C. Pflugl, L. Diehl, *et al.*, "1.6W high wallplug efficiency, continuous-wave, room temperature quantum cascade laser emitting at 4.6 μm," *Appl. Phys. Lett.* **92**, 111110 (2008).
4. T. Gresch, J. Faist, and M. Giovannini, "Gain measurements in strain-compensated quantum cascade laser," *Appl. Phys. Lett.* **94**, 161114 (2009).
5. J. C. Shin, M. D'Souza, Z. Liu. *et al.*, "Highly temperature insensitive, deep-well 4.8 μm emitting quantum cascade semiconductor lasers," *Appl. Phys. Lett.* **94**, 201103 (2009).
6. J. C. Shin, L. J. Mawst, D. Botez, I. Vurgaftman, and J. R. Meyer, "Ultra-low temperature sensitive, deep-well quantum-cascade lasers (λ = 4.8 μm) via uptapering the conduction band edge of injector regions," *Electronics Lett.* **45**, 741 (2009).
7. U. Bockelmann and G. Bastard, "Phonon scattering and energy relaxation in two-, one-, and zero-dimensional electron gases," *Phys. Rev. B* **42**, 8947 (1990).

8. H. Benisty, C.M. Sotomayor-Torres, and C. Weisbuch, "Intrinsic mechanism for the poor luminescence properties of quantum-box systems," *Phys. Rev. B* **44**, 10945 (1991).
9. R. Heitz, M. Grundmann, N. N. Ledentsov, *et al.*, "Multiphonon-relaxation processes in self-organized InAs/GaAs quantum dots," *Appl. Phys. Lett.* **68**, 361 (1996).
10. J. Urayama, T. B. Norris, J. Singh, and P. Bhattacharya, "Observation of phonon bottleneck in quantum dot electronic relaxation," *Phys. Rev. Lett.* **86**, 4930 (2001).
11. L. Rebohle, F. F. Schrey, S. Hofer *et al.*, "Energy level engineering in InAs quantum dot nanostructures," *Appl. Phys. Lett.* **81**, 2079 (2002).
12. S. Sauvage, P. Boucaud, R. P. Lobo, *et al.*, "Long polaron lifetime in InAs-GaAs self-assembled quantum dots," *Phys. Rev. Lett.* **88**, 177402 (2002).
13. E. A. Zibik, L. R. Wilson, R. P. Green, *et al.*, "Intraband relaxation via polaron decay in InAs self-assembled quantum dots," *Phys. Rev. B* **70**, 161305 (2004).
14. B. Kochman, A. D. Stiff-Roberts, S. Chakrabarti, *et al.*, "Absorption, carrier lifetime, and gain in InAs-GaAs quantum-dot infrared photodetectors," *IEEE J. Quantum Electronics* **39**, 459 (2003).
15. L. Jacak, J. Krasnyi, D. Jacak, and P. Machnikoski, "Magnetopolaron in a weakly elliptical InAs/GaAs quantum dot," *Phys. Rev. B* **67**, 035303 (2003).
16. C.-F. Hsu, J.-S. O, P. S. Zory, and D. Botez, "Intersubband laser design using a quantum box array," *Proc. SPIE* **3001**, 271 (1997).
17. C.-F. Hsu, J.-S. O, P. S. Zory and D. Botez, "Intersubband quantum-box semiconductor lasers," *IEEE J. Selected Topics Quantum Electronics* **6**, 491(2000).
18. I. A. Dmitriev and R. A. Suris, "Quantum cascade lasers based on quantum dot superlattice," *Physica status sol. A* **202**, 987 (2005).
19. L. Diehl, D. Bour, S. Corzine, *et al.*, "High-temperature continuous wave operation of strain-balanced quantum cascade lasers grown by metal organic vapor-phase epitaxy," *Appl. Phys. Lett.* **89**, 081101 (2006).
20. I. C. Sandall, P. M. Smowton, C. L. Walker, H. Y. Liu, M. Hopkinson, and D. J. Mowbray, "Recombination mechanisms in 1.3 μm InAs quantum-dot lasers," *IEEE Photon. Technol. Lett.* **18**, 965 (2006).
21. D. Botez, G. Tsvid, M. D'Souza, *et al.*, "Progress towards intersubband quantum-box lasers for highly efficient cw operation in the mid-infrared," *J. Nanophotonics* **3**, 031606 (2009).
22. D. Botez, L. J. Mawst, G. Peterson, and T. J. Roth, "Resonant optical transmission and coupling in phase-locked diode-laser arrays of antiguides: The resonant-optical-waveguide array," *Appl. Phys. Lett.* **54**, 2183 (1989).
23. H. Yang, L. J. Mawst, M. Nesnidal, J. Lopez, A. Bhattacharya, and D. Botez, "10W near-diffraction-limited pulsed power from 0.98 μm-emitting, Al-free phase-locked antiguided arrays," *Electronics Lett.* **33**, 136 (1997).
24. H. Yang, L. J. Mawst, and D. Botez, "1.6 W continuous-wave coherent power from large-index-step, resonant, antiguided diode laser arrays," *Appl. Phys. Lett.* **76**, 1219 (2000).

25. S. Li, D. Xu, and D. Botez, "High power, single-mode operation from photonic-lattice semiconductor lasers with controllable lateral resonance," *Appl. Phys. Lett.* **88**, 091112 (2006).
26. S. Xiao, X. Yang, E. W. Edwards, Y.-H. La, and P. F. Nealey, "Graphoepitaxy of cylinder-forming block copolymers for use as templates to pattern magnetic metal dot arrays," *Nanotechnol.* **16**, S324 (2005).
27. C. Zmudzinski, D. Botez, L. J. Mawst, C. Tu, and L. Frantz, "Coherent, one-watt cw operation of large aperture resonant arrays of antiguided diode lasers," *Appl. Phys. Lett.* **62**, 2914 (1993).
28. J. R. Leger, G. J. Swanson, and M. Holz, "Efficient side-lobe suppression of laser diode arrays," *Appl. Phys. Lett.* **50**, 1044 (1987).
29. M. K. Rathi, G. Tsvid, A. A. Khandekar, J.C. Shin, D. Botez, and T. F. Kuech, "Passivation of interfacial states for GaAs- and InGaAs/InP-based regrown nanostructures," *J. Electronic Mater.* **38**, 2023 (2009).

GaSb-based Type-I Laser Diodes Operating at 3 μm and Beyond

L. Shterengas, G. Kipshidze, T. Hosoda, and G. Belenky
Department of Electrical and Computer Engineering, SUNY–Stony Brook
Stony Brook, NY 11794-2350, U.S.A.

1. Introduction

Laser light sources operating in the spectral region near 3 μm are required for variety of applications, ranging from aesthetic surgery to remote spectroscopic analysis. Semiconductor diode lasers operating at room temperature (RT) with watt-level continuous wave (cw) output power can benefit the design of the corresponding systems due to their inherent stability, compactness, and possibly low cost and high efficiency. Until recently, the very idea of cw diode lasers operating at RT with wavelength $\lambda \sim 3$ μm was thought to be questionable. This pessimism originated in the well-known fundamental increases of the nonradiative Auger recombination and of free carrier absorption with wavelength. The associated carrier and photon losses were considered likely to prevent the mid-infrared (mid-IR) diode lasers from reaching cw lasing at RT.

Today, cw RT operation of 3 μm diode lasers with hundreds of mW of output power is a reality. Clearly, this experimental fact casts doubt on the validity of the old argument based on the destructive power of Auger recombination and free carrier absorption. Moreover, the narrow bandgap nature of the active region not only brings disadvantages but also can provide significant benefits. The advantage of using narrow bandgap material for generation of optical gain is rather straightforward, though rarely discussed in context of mid-IR laser development. The electron density of states in the Γ-valley is nearly proportional to the direct bandgap in III-V binary materials.[1] This proportionality is more or less preserved in corresponding ternary and quaternary alloys used in quantum wells (QWs) of diode lasers. In QWs compressively strained above 1%, the top valence subband has an in-plane effective mass that is rather close to electron effective mass.[2] The resulting reduced density of states facilitates separating the quasi-Fermi levels in narrow bandgap active QWs under injection. Since bulk transition matrix element does not show pronounced material bandgap dependence,[1] this leads to improved differential gain in narrow bandgap QWs. Hence, it is expected that mid-IR diode lasers can have much lower transparency and, possibly, lower threshold carrier concentrations as compared to their near-IR counterparts. Auger recombination is a three-particle process and its net rate is superlinear in carrier concentration. Hence, the threshold current density of the mid-IR diode lasers can be rather low despite the increased probability of individual Auger events in narrow bandgap

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

QWs. It is worth mentioning that the spontaneous recombination probability decreases dramatically with the bandgap, thus furnishing another advantage for narrow bandgap QWs.

The temperature stability of mid-IR lasing thresholds remains an open question because Auger rate can be highly temperature sensitive. Hence, even if the laser threshold current is relatively low, as long as it is determined by Auger processes it could increase dramatically at elevated temperatures. However, it has been argued that, under certain conditions, the Auger rate in QWs can have reduced temperature sensitivity.[3] Identification of the effects of Auger recombination on laser parameters is complicated by the experimental difficulties in distinguishing between penalties associated with Auger recombination and thermal carrier delocalization.

In this chapter, we demonstrate that GaSb-based diode lasers with improved carrier confinement in type-I InGaAsSb QWs can operate at room temperature up to $\lambda = 3.5\ \mu m$. In Section 2 we will describe the corresponding laser technology and review the material development for fabrication of the diode lasers operating near and above 3 μm. Section 3 will present an example of design optimization aimed at improving the output power of the 3 μm emitters. The conclusion will summarize our experimental observations and discuss future prospects.

2. Diode lasers operating at RT in the 3–3.5 μm spectral range

Our laser heterostructures were grown by solid-source molecular beam epitaxy using VEECO GEN-930 reactor on tellurium doped GaSb substrates. The cladding layers were lattice-matched to GaSb quaternary AlGaAsSb alloys with Al content near 85%. Beryllium and tellurium were used for *p*- and *n*-type doping, respectively. The laser active region contained several compressively-strained InGaAsSb QWs with In content in the vicinity of 50%. The utilization of the compressive strain was previously identified as an effective way to improve hole confinement in QWs.[2] The barrier and waveguide material was the quinary AlInGaAsSb alloy lattice-matched to GaSb. The use of the quinary In-containing material as QW barriers also improves hole confinement.[4,5] The results presented in Refs. 2, 4, and 5 conclusively demonstrate that enhancement of the hole confinement barrier is crucial factor in reducing the threshold current density of GaSb-based diode lasers at room temperature.

Wafers were processed into 100 μm wide index-guided ridge lasers by wet etching of the top-cladding layer outside of the current stripe. Laser mirrors were cleaved to form I and 2 mm-long cavities. For cw laser parameter characterization the devices were mirror-coated with reflectivity below 5% for anti-reflection (AR) and above 90% for high reflection (HR) coatings. Coated laser dies were soldered epi-side down with indium on polished gold-coated copper blocks.

Figure 1(a) shows the cw light-current characteristic of diode lasers emitting at 3 μm. These devices demonstrated cw threshold current densities of about 280

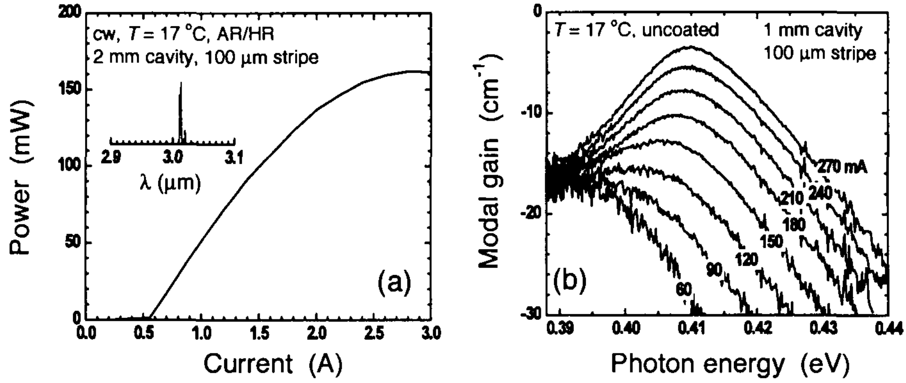


Figure 1. (a) Light-current characteristics measured in cw regime at 17 ^{0}C for 2 mm long, 100 μm wide, AR/HR coated diode lasers emitting near 3 μm; (b) current dependence of the modal gain spectra for corresponding 1 mm long, 100 μm wide, uncoated lasers measured in pulsed regime (200 ns/2 MHz).

A/cm^2 and produced more than 150 mW of output power. The laser heterostructure had two 12 nm wide InGaAsSb QWs spaced 40 nm apart, with compressive strain about 1.7%. The two-QW active region was placed in the center of waveguide region made of quinary $In_{0.2}Al_{0.2}Ga_{0.6}As_{0.19}Sb_{0.81}$ material. The total waveguide core width (from *p*-cladding to *n*-cladding) was about 1070 nm.

The modal gain was measured by Hakki-Paoli method for 1 mm long uncoated devices – see Fig. 1(b). The estimated internal optical loss is ~5 cm^{-1} and transparency current density is about 100 A/cm^2. The relatively low values of threshold current density and internal optical loss indicate that there is no dramatic increase of nonradiative recombination and free carrier absorption losses, at least for λ near 3 μm. The band offsets between QW and quinary barrier materials were estimated as about 300 meV and 160 meV in conduction and valence bands, respectively.[1,6]

Figure 2(a) plots the cw light-current characteristic of diode lasers emitting near 3.17 μm at room temperature. These devices demonstrated cw threshold current densities of about 480 A/cm^2 and produced more than 50 mW of the output power. The laser heterostructure had four 14 nm wide InGaAsSb QWs spaced 40 nm apart, with compressive strain about 1.6%. The number of QWs in active region was twice that of 3 μm emitting devices with provision to decrease the threshold carrier concentration. The four-QW active region was placed in the center of waveguide region made of quinary $In_{0.2}Al_{0.2}Ga_{0.6}As_{0.19}Sb_{0.81}$ material. The total waveguide core width (from *p*-cladding to *n*-cladding) was about 1180 nm.

The current dependence of the modal gain spectrum measured by Hakki-Paoli method for 1 mm long uncoated devices is shown in Fig. 2(b). The estimated

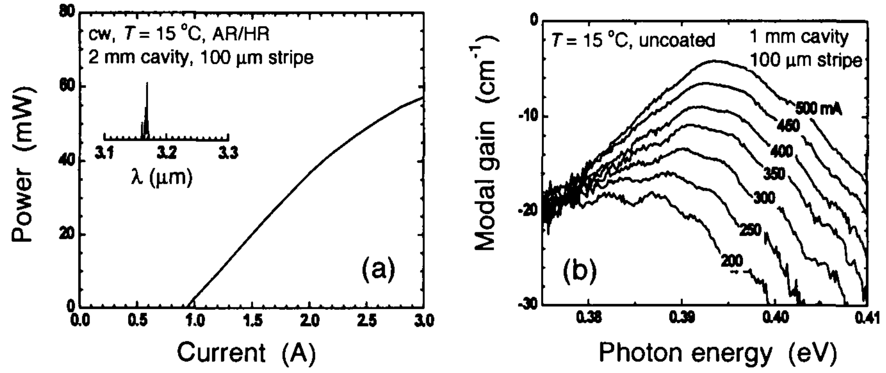


Figure 2. (a) Light-current characteristics measured in cw regime at 15 ^{0}C for 2 mm long, 100 μm wide, AR/HR coated diode lasers emitting near 3.17 μm; (b) current dependence of the modal gain spectra for corresponding 1 mm long, 100 μm wide, uncoated lasers measured in pulsed regime (200 ns/2 MHz).

internal optical loss is 8–10 cm^{-1} and transparency current density is about 200 A/cm^2. The twofold increase in the transparency current density compared to the 3 μm emitters of Fig. 1(b) is normal, given that the active region contains twice the number of QWs. The apparent reduction of laser efficiency and increase in the threshold of 3.17 μm emitters as compared to 3 μm emitters is partly due to weaker hole confinement. The valence band offsets between QWs and quinary barrier materials at the longer wavelength are smaller by ~40 meV.

Increasing of the lasing wavelength above 3.17 μm is expected to degrade the hole confinement even further. Devices operated in cw at RT with λ = 3.36 μm produced only 15 mW of power.[7] These laser heterostructure contained four 16 nm wide QWs in the active region of a waveguide of ~1150 nm total width. Internal optical losses were estimated to be about 12–15 cm^{-1}. Threshold current density was about 750 A/cm^2 at 17 ^{0}C, for 2 mm long, 100 μm wide, AR/HR coated etched lasers.

Figure 3(a) plots the pulsed light-current characteristics of diode lasers designed to operate above 3.5 μm at RT. The laser heterostructure had four 15 nm wide InGaAsSb QWs spaced 40 nm apart, with compressive strain of ~1.5%. The four-QW active region was placed in the center a quinary $In_{0.25}Al_{0.2}Ga_{0.55}$ $As_{0.23}Sb_{0.77}$ waveguide. The total waveguide core width (from *p*-cladding to *n*-cladding) was about 700 nm. The lasers were mirror-coated to reflect 30% (NR) and above 90% (HR).

A dramatic increase in the pulsed threshold current density up to above 1000 A/cm^2 at 280 K prevented the lasers from operating in cw regime at room temperature – see Fig. 3(b).

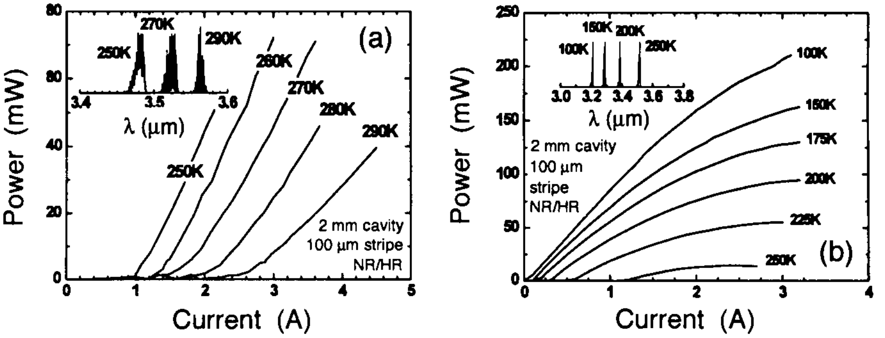


Figure 3. (a) Pulsed (200 ns/100 kHz) and (b) cw light-current characteristics measured in the 100–290 K temperature range for 2 mm long, 100 μm wide, NR/HR coated lasers.

Clearly, the internal optical losses increase with laser wavelength from about 5 cm^{-1} at 3 μm up to 12–15 cm^{-1} at 3.36 μm. A part of this increase is due to the increased overlap of the optical field with doped cladding regions from about 3–4% at 3 μm to 4–5% at 3.4 μm (for waveguide width of ~1100 nm). The rest of the internal loss increase can be accounted for by the rise of the free hole absorption coefficient with λ,[8] as well as enhanced free hole absorption in the active region due to a greater number of QWs and less-localized holes. A substantial reduction of the internal optical losses can be achieved by either minimizing the overlap of the optical field with *p*-doped regions or by minimizing the doping level (or both). Strong carrier localization in the QWs is expected to reduce the concentration of free holes in waveguide region, also reducing the net internal optical losses.

The threefold increase in internal optical loss from 5 cm^{-1} at 3 μm to 12–15 cm^{-1} at 3.36 μm corresponds to a less than twofold increase of the total loss for 2 mm long AR/HR coated devices (mirror loss is ~8 cm^{-1}) and hence cannot explain the observed increase of the threshold current density. Indeed, the differential gain as a function of the current, as determined from Hakki-Paoli measurements shown in Figs. 1(b) and 2(b), decreases with wavelength by 40% and 55% when wavelength is changed from 3 to 3.17 μm and from 3 to 3.36 μm, respectively. The estimated reduction of the MQW optical confinement factor between 3 μm emitters and 3.36 μm emitters is below 10%; hence contribution of this effect to observed reduction of the differential gain is marginal.

The gradual decrease in the differential gain can be accounted for by the degradation of hole confinement, degradation of injection efficiency, and enhanced Auger recombination. It is likely that these effects are mutually reinforcing. For instance, reduced hole confinement leads to the thermal population of the waveguide valence band that, in turn, can reduce injection efficiency. Additionally, thermal population of the barrier regions in the valence band effectively "increases the hole density of states" in QWs, thus increasing the threshold carrier concentration and enhancing Auger recombination.

3. Diode lasers operating at 3 μm with 300 mW cw RT output power

Here we discuss the optimization of the diode laser heterostructure by modifying a single parameter – the waveguide width. The optimization of the waveguide width was performed for 3 μm emitters and led to twofold increase of the cw RT output power level, *i.e.* from 160 mW of Fig. 1(a) to more than 300 mW.

Three laser heterostructures were grown with different waveguide widths: (a) 470 nm, (b) 1070 nm (Fig. 1), and (c) 1470 nm. The heterostructure parameters other than waveguide width were nominally kept the same for all the three lasers. Figure 4(a) shows pulsed (200 ns, 10 kHz) light-current characteristics measured at room temperature for 1 mm long uncoated lasers. The light-current characteristics demonstrate a rollover for all three structures that becomes more severe with increasing the waveguide width. Figure 4(b) shows the threshold current densities measured for three structures over a wide temperature range. At 150 K, all three devices have a very low threshold current density of ~25 A/cm^2 that increases with temperature at a rate that depends on waveguide width. At 300 K the threshold current density is more than two times higher for devices with the widest waveguide than for the narrowest one.

The earlier rollover of the light-current characteristics in devices with broader waveguides can be accounted for by the carrier transport constraints in waveguide region.[9] Accumulation of free carriers in the waveguide between the cladding and the two-QW active region can lead to reduction of the slope efficiency due to both free carrier absorption and recombination.

An analysis of the dependence of the threshold current density on waveguide width shown in Fig. 4(b) should take into account the optical field distribution in the heterostructure. The optical field distribution leads to 1.7 times higher MQW optical confinement factor in lasers with 470 nm waveguides then in devices with 1470 nm ones. Enhancement of the optical confinement factor should improve the differential gain in narrow waveguide lasers. Figure 5(a) shows the experimental

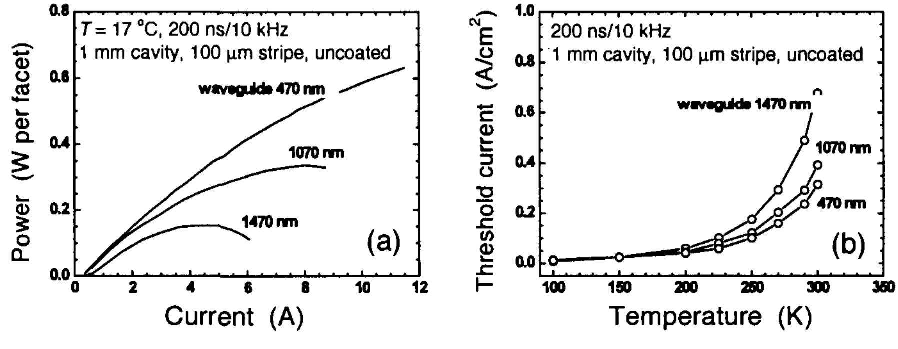


Figure 4. (a) Pulsed (200 ns/10 kHz) light-current characteristics and (b) temperature dependence of the threshold current density for 1 mm long, 100 μm wide, uncoated 3 μm lasers for different waveguide widths.

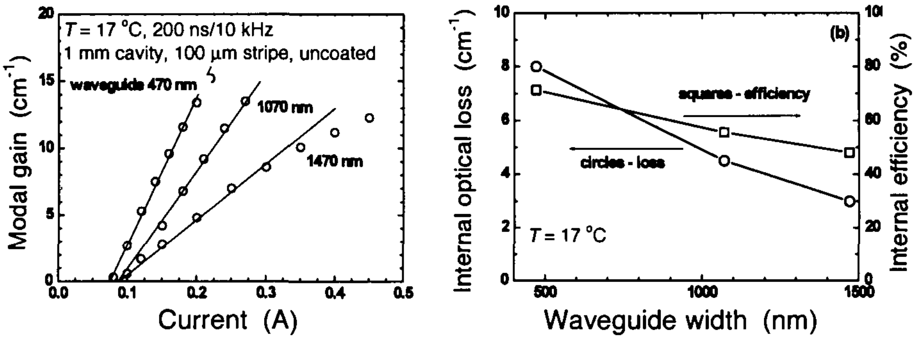


Figure 5. (a) Peak modal gain *vs.* current for 3 μm lasers with different waveguide widths; (b) internal optical loss and internal efficiency *vs.* waveguide width.

dependence of peak modal gain on the current, as measured by Hakki-Paoli method for all lasers. The differential gain of the devices with the narrowest waveguide is about 2.5 times higher than in lasers with the widest waveguide, *i.e.* the experimental rate of the differential gain increase exceeds the improvement in optical confinement.

Thus, differential gain is enhanced by improved device injection efficiency, as shown in Fig. 5(b). Since the diffusion current through a narrower waveguide can be supported by a lower carrier concentration, we expect the waveguide recombination current to be smaller in narrower waveguides. Figure 5(b) shows that the broadening of the waveguide decreases internal losses, as expected.[10] However, the combined effects of the carrier transport through the waveguide and poorer MQW optical confinement factor did not allow the 3 μm lasers to benefit from associated internal optical loss reduction.

Based on the experimental studies of the effect of the waveguide width on laser parameters, we optimized the 3 μm laser heterostructure. Diode lasers with 570 nm waveguide width and reduced *p*-cladding doping (10^{17} cm^{-3} instead of the 2×10^{17} cm^{-3} used previously) were grown to minimize optical loss while maximizing QW optical confinement and internal efficiency. Figure 6(a) shows cw RT light-current and power conversion characteristics of 2 mm long, 100 μm wide, AR/HR coated devices. The maximum power of 310 mW was achieved at a current of 4.7 A at cooling water temperature of 17 °C. The power conversion efficiency was better than 5% for the whole range of operation with maximum value approaching 8% at ~150 mW output power. The cw threshold current density was 200 A/cm^2 (100 A/cm^2 per QW). The voltage drop across laser heterostructure was below 1.4 V at the maximum output power level.

The current dependence of the modal gain spectrum measured by Hakki-Paoli method for 1 mm long uncoated devices is shown in Fig. 6(b). The estimated internal optical loss is 4–5 cm^{-1} and transparency current density is below 100 A/cm^2. Differential gain as a function of current is better than in previous devices of Fig. 1(b). The external efficiency near threshold is 0.25 photons per electron, *i.e.* still only quarter of the cw current above threshold was being converted into light. This value should increase once the internal optical losses are minimized and carrier delivery to active QWs is further improved.

4. Conclusions

As of the time of writing, modern GaSb-based type-I QW diode laser technology can produce diode lasers operating in cw regime at room temperature up to 3.36 μm;[7] similar designs can emit above 3.5 μm in pulsed mode. The major breakthrough in mid-IR GaSb-based diode laser performance parameters was achieved when the hole confinement in active QWs was augmented. The optimization of the laser heterostructure design can effectively improve the device differential gain and internal efficiency. Diode lasers operating near 3 μm can now produce above 300 mW in cw regime with power conversion efficiency near 8%.

High differential gain and relatively low internal optical losses have reduced the threshold current density of 3 μm GaSb-based emitters (200 A/cm^2) to values comparable with the best ~1 μm GaAs-based high power diode lasers (50 A/cm^2).[11] It is worth noting that the internal efficiency of those GaAs-based state-of-the-art devices is approaching 100%, while GaSb-based 3 μm emitters have internal efficiency near 50%. Nearly twofold difference in internal efficiencies implies even closer correspondence between actual QW threshold current densities of near- and mid-IR devices. This result implies that the "Auger limit" does not curtail the performance of mid-IR diode lasers, at least in spectral region near 3 μm.

We speculate that the output power of the best current GaSb-based diode lasers is limited by the carrier transport constraints. Once the carrier delivery into active QWs is improved, the output power of lasers operating in the 3–3.5 μm spectral region is expected to increase.

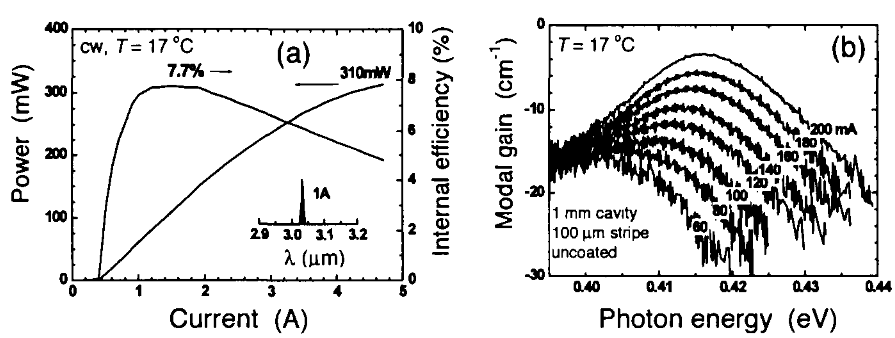


Figure 6. (a) Light-current characteristics measured in cw regime at 17 °C for 2 mm long, 100 μm wide, AR/HR coated optimized 3 μm diode lasers; (b) modal gain spectra *vs.* current for corresponding 1 mm long, 100 μm wide, uncoated lasers measured in pulsed regime (200 ns/2 MHz).

Acknowledgment

This work was supported by the US Air Force Office of Scientific Research under contract FA95500810458 and US Army Research Office under contract W911NF0610399.

References

1. I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, "Band parameters for III-V compound semiconductors and their alloys," *J. Appl. Phys.* **89**, 5815 (2001).
2. J. Chen, D. Donetsky, L. Shterengas, M. Kisin, G. Kipshidze, and G. Belenky, "Effect of quantum well compressive strain above 1% on differential gain and threshold current density in type-I GaSb-based diode lasers," *IEEE J. Quantum Electronics* **44**, 1204 (2008).
3. A. S. Polkovnikov and G. G. Zegrya, "Auger recombination in semiconductor quantum wells," *Phys. Rev. B* **58**, 4039 (1998).
4. M. Grau, C. Lin, O. Dier, C. Lauer, and M.-C. Amann, "Room-temperature operation of 3.26 μm GaSb-based type-I lasers with quinternary AlGaInAsSb barriers," *Appl. Phys. Lett.* **87**, 241104 (2005).
5. T. Hosoda, G. Belenky, L. Shterengas, G. Kipshidze, and M. Kisin, "Continuous-wave room temperature operated 3.0 μm type I GaSb-based lasers with quinternary AlInGaAsSb barriers," *Appl. Phys. Lett.* **92**, 091106 (2008).
6. G. P. Donati, R. Kaspi, and K. J. Malloy, "Interpolating semiconductor alloy parameters: Application to quaternary III-V band gaps," *J. Appl. Phys.* **94**, 5814 (2003).
7. L. Shterengas, G. Belenky, T. Hosoda, G. Kipshidze, and S. Suchalkin, "Continuous wave operation of diode lasers at 3.36 μm at 12 degrees C," *Appl. Phys. Lett.* **93**, 011103 (2008).
8. A. Chandola, R. Pino, and P. S. Dutta, "Below bandgap optical absorption in tellurium-doped GaSb," *Semicond. Sci. Technol.* **20**, 886 (2005).
9. B. S. Ryvkin and E. A. Avrutin, "Asymmetric, nonbroadened large optical cavity waveguide structures for high-power long-wavelength semiconductor lasers," *J. Appl. Phys.* **97**, 123103 (2005).
10. D. Botez, "Design considerations and analytical approximations for high continuous-wave power, broad-waveguide diode lasers," *Appl. Phys. Lett.* **74**, 3102 (1999).
11. D. Fekete, M. Yasin, A. Rudra, and E. Kapon, "Very low transparency currents in double quantum well InGaAs semiconductor lasers with delta-doped resonant tunneling," *Appl. Phys. Lett.* **92**, 021109 (2008).

This page intentionally left blank

Bridging Optics and Electronics with Quantum Cascade Lasers, Antennas, and Circuits

J. Faist, C. Walther, M. Amanti, G. Scalari, M. Fischer, and M. Beck
Dept. of Physics, ETH Zurich, 8083 Zurich, Switzerland

1. Introduction

The terahertz frequency range, loosely defined as the range between 1–10 THz, has traditionally been lacking in convenient optical sources. Terahertz radiation has some important potential applications in imaging, chemical sensing and telecommunications.[1,2] The long wavelength of the terahertz radiation enables a good penetration into heavily diffusing materials, such as foams or small-grained material. The vibrational modes of large organic and inorganic molecules are in the THz region, offering potential avenues for label-free sensing of DNA.[3] The THz range has also some very interesting radio astronomy applications, as it contains, among other important spectral signatures, the brightest cooling lines of the dense interstellar medium ([CII] 158 μm, [OI] 63 and 145 μm) present in space. Because of the lack of transparency of the atmosphere in the THz region, such observation must be performed from airborne or spaceborne telescope such as SOPHIA, putting additional constrains on optical sources. For this reason, there is a strong interest in the development of terahertz sources that are tunable, portable and manufacturable in large quantities. Recently, much progress has been made in sources based on nonlinear generation from either ultrafast lasers or continuous wave near-infrared or even mid-infrared sources.[4] Compared to these techniques based on nonlinear mixing, fundamental oscillators have the potential advantage of offering higher efficiency and an easier fabrication technology.

The operation of quantum cascade lasers (QCLs)[5,6] was first achieved in the THz in 2001 at a frequency of 4.1 THz, in pulsed mode and up to a maximum operating temperature of 65 K.[7] Continuous wave operation was achieved shortly thereafter,[8] followed by operation above liquid nitrogen temperatures.[9] The introduction of a design based on resonant optical phonon extraction,[10] combined with the use of metal-metal waveguides, has enabled a significant increase in the maximum operating temperature achieved by these devices, up to a maximum temperature of 186 K.[11] Even if this last result is a significant improvement compared to the first generation devices, it still lies significantly below the temperatures achievable by Peltier thermoelectric coolers. Another challenge was the extension of the operating range beyond the initially demonstrated 3–4 THz. On the low frequency side, optimization of the design of a bound-to-continuum transition combined with an electron reservoir energetically decoupled from the miniband enabled operation of these devices down to a frequency of 1.2 Thz.[12,13]

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

2. Four-quantum-well active region

As shown schematically in Fig. 1, the active stage of the device presented here is based on a bound-to-continuum transition, while the miniband is coupled to a extractor quantum well.[14] This architecture has the advantage of reducing the direct coupling between the upper state and the extractor well because they are physically separated by the length of the miniband region. The alignment condition on the extraction stage is also somewhat relaxed by the fact that it has to be satisfied over the width of the miniband rather than of a single state. The original design following this architecture[14] was based on a five-quantum-well active region and a two-quantum-well injector region. We report here the study of a shortened version of this design, where the total number of quantum wells has been reduced to four, with a three-quantum-well active region coupled to a single quantum well injector. The latter choice was made to increase the maximum operating temperature of the device.

A schematic band diagram of the structure, computed using a self-consistent Schrödinger–Poisson solver is shown in Fig. 1. Shown also are the energies and electron probability densities, computed in the extended basis spanning the whole period length. The oscillator strength is concentrated mainly on the transition

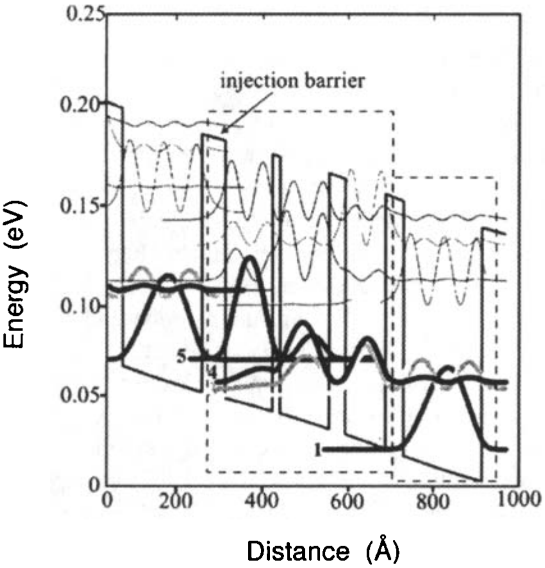


Figure 1. Schematic conduction band diagram of one stage of the structure EV1116 under an applied electric field of 7.6×10^3 V/cm. The moduli squared of the relevant wavefunctions are shown. The layer sequence of one period of the $Al_{0.15}Ga_{0.85}As/GaAs$ structure, starting from the injection barrier is **5.5**/11.0/**1.8**/11.5/**3.8**/9.4/**4.2**/18.4 in nm, where the $Al_{0.15}Ga_{0.85}As$ layers are in bold. The 18.4 nm GaAs layer is doped with Si to $N_D = 2\times10^{16}$ cm^{-3}.

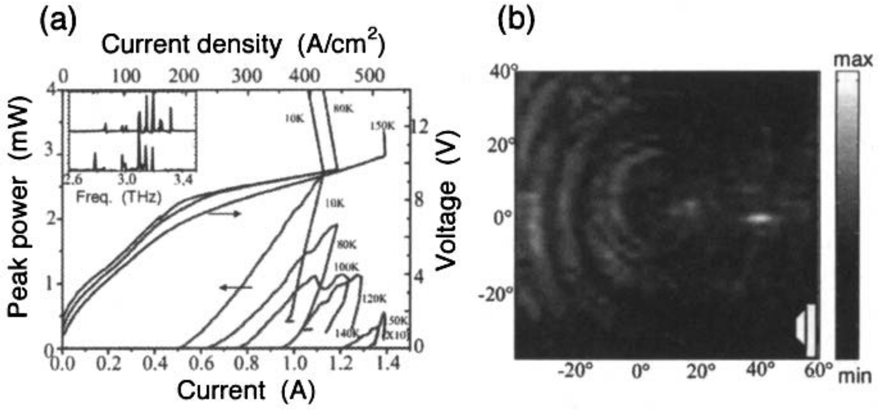


Figure 2. (a) Light and bias voltage *vs.* current of a four-quantum-well active region device measured a various temperatures, as indicated. The maximum operating temperature reached by this device is 150 K. Inset shows the spectrum of device measured for two different drive currents. (b) Far-field spectrum of the device: horizontal axis is the growth direction, vertical axis is in the plane of the device.

between levels 5 and 4 of the active region. As shown in Fig. 1, the wavefunctions of these two states have a relatively weak spatial overlap, that translate into a reduced normalized oscillator strength $f' = (m^*/m_0)\cdot f = 0.31$. As expected for such a diagonal transition, the optical phonon scattering rate yields a relatively long lifetime at elevated temperature ($\tau = 1.9$ ps at 150 K).

The light–current characteristics of the MBE-grown sample EV1136 is reported as a function of temperature in Fig. 2. The maximum temperature achieved in the device shown is 158 K.

3. Antenna-like third-order distributed feedback QCL

The problems with the devices based on metal-metal waveguides, like the one shown in Fig. 2, consist of both relatively low output power as well as a very broad far-field emission. These problems arise from the use of a metal-metal waveguide and they severely limit the usefulness of the device for applications. To try and alleviate these problems, we investigated a distributed feedback (DFB) grating[15] resonant with the third-order Bragg condition.

In our approach, in contrast to approaches based on surface emission,[16] we optimize the natural edge emission of the TM-polarized light in the structure by utilizing a third-order DFB grating. This simultaneously provides the feedback for the optical mode inside the semiconductor and the necessary diffraction to couple the radiation out to free space. An intuitive picture of the system is given in Fig. 3(a): a waveguide mode with propagation constant $\beta = n_{\mathrm{eff}}k_0$ (black arrow) incident on an infinite length grating with periodicity Λ, is diffracted back by the reciprocal vectors $G = 2\pi/\Lambda$ (light arrows). In the case of a third-order grating $G = 2\beta/3$, and when the refractive index $n_{\mathrm{eff}} = 3$, the first two diffractions match the wavevector of the mode propagating in the air in the horizontal direction, while the third one provides the feedback in the waveguide. The estimated effective refractive index for double metal waveguides is ~3.5, a value close to the relation $n_{\mathrm{eff}} = 3$ that would allow the use of the grating to extract the light out of the semiconductor. Efficient coupling is nevertheless achieved because of the finite length of the large refractive index contrast grating.

The latter has been realized by deep dry etching of the semiconductor along the waveguide, alternating active region and air with a duty cycle of ~10%, as shown in Fig. 3(b). To avoid the parasitic higher-order lateral and longitudinal modes, we used the same geometry reported for the photonic crystal structures of reference:[17] all around the waveguide we left an unmetallized region, in order to use the thin heavily-doped semiconductor lying on the top of the waveguide as an absorbing layer.

We designed the grating to overlap the gain frequency of our active medium with the mode on the left side of the gap, where the electromagnetic energy is more confined in the semiconductor region.

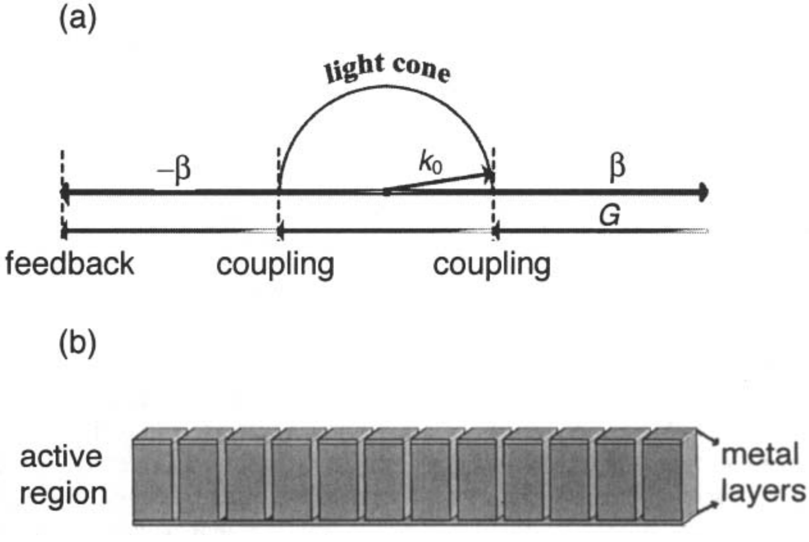


Figure 3. (a) Vector diagram for a third-order DFB in the wavevector space k. Arrows labeled with β represent modes propagating inside the semiconductor and arrows labeled with G represent the grating diffraction vectors. The upper half circle has a radius equal to the wavevector of the mode propagating in the air k_0, so it represents the light cone: vectors ending on the circle can couple out of the semiconductor. (b) Schematic drawing of the active region sandwiched in the metal-metal waveguide with the deep-etched third-order grating.

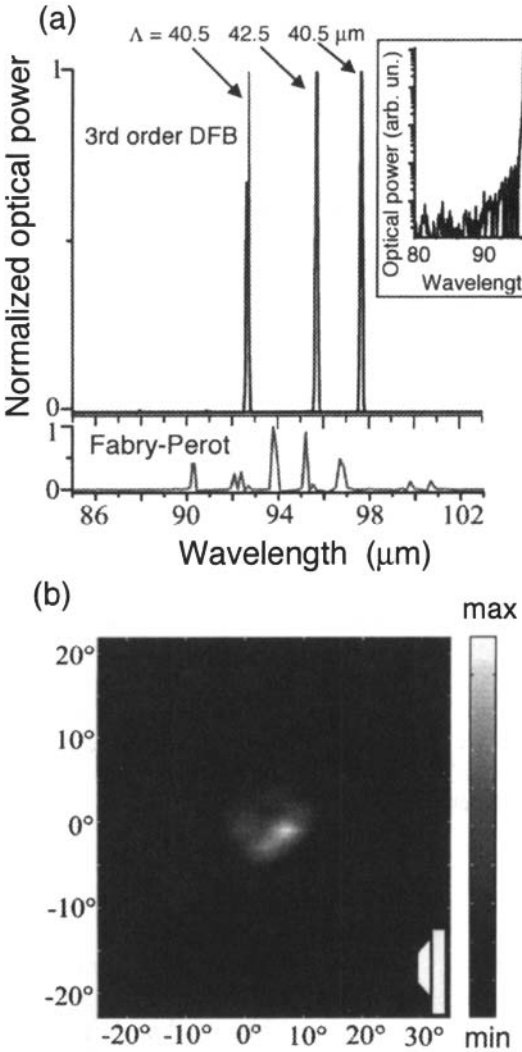


Figure 4. (a) Normalized emission spectra of the third-order DFB for different grating periods Λ. Inset: spectrum in logarithmic scale. (b) Far-field of the emitted spectrum. The optical power reaches 14 mW in pulsed mode.

For all the devices we measured the light–current–voltage (*L*–*I*–*V*) curves in pulsed mode at 10 K and the emission spectra for different operating currents with a resolution of 0.1 cm^{-1}. The output peak power measured for a typical DFB device was 15 mW, a factor of 7 higher than for a standard double-metal waveguide of similar dimensions. The slope efficiency was also improved, reaching the value of 130 mW/A, considerably higher than the highest value of 40 mW/A shown for surface emitting structures.[16,18] This is due to the combined effects of improved extraction and collection efficiency. All the devices worked in single mode operation with 30 dB sideband suppression ratio; their emission frequency can be tuned over 0.17 THz by changing the grating period – see Fig. 4(a). These two results, the enhancement of the output power and of the slope efficiency, already prove the simple model given above of the grating as means to outcouple single mode laser light from a double metal waveguide. Moreover, the laser performance of our structure has another important advantage: the reshaping of the emission pattern in to a single lobed beam. To understand the effect of the grating, we can turn to the well-known concept of a linear phased array in antenna theory: the periodic apertures define a series of sources phase-shifted by π along the waveguide. Constructive interference from the different emitters can then be observed only in a narrow spot emitted in the direction along the array.

A systematic characterization of the far-field laser beam pattern has been carried out on the devices. The results are presented on two-dimensional graphs where the α and β axes correspond respectively to angular scans in the planes perpendicular to and parallel to the grating slits: point (0°, 0°) corresponds to a longitudinal edge emission, whereas point (90°, 0°) corresponds to surface emission. Data were acquired over a range of 150° along the horizontal direction, but for more clarity we show a magnification in the region where the beam is observable.

In Fig. 4(b) we show the measured far field for a typical DFB structure: the laser light in emitted in a single-lobed narrow beam of ~10° of divergence. We show the control of the laser emission wavelength and the enhancement of the output power due to the improvement in extraction efficiency. Moreover, we show that the grating can act as an array of phased linear sources, reshaping the typical wide and patterned far field of double-metal waveguides into a narrow beam of ~10° divergence in both directions.

4. Circuit-based QCLs

In a number of applications, it would be desirable to have lasers with a very small active volume. Dielectric-based resonators are fundamentally limited to a volume of the order of $(\lambda/n)^3$. In contrast, metallic-based cavities do not suffer from the same kind of limitations. In fact, resonant inductance capacitor (LC) circuits can confine the electric field in an almost arbitrarily small volume. We investigate such cavities in which the dielectric of the capacitor is replaced by a QCL material.

Besides the small volume, another interesting property of such cavities is that they are inherently single-mode, with a well-behaved far field. Control of the emission frequency is trivial and can be achieved by merely changing the value of capacitance or inductance. Because of their small size and the fact that their frequency can easily be set, such devices lend themselves easily to the fabrication of arrays that could cover a wide spectral band. The radiative Q of such cavity will grow as λ^3/V, *i.e.* the quality factor will increase as the volume V decreases. As noted by many recent theoretical contributions, the characteristics of such a "small laser" would depart significantly from the predictions of semi-classical theory and exhibit new features not observable in conventional structures.[19]

An attractive feature of QCL materials is the wide bandwidth of the gain material. The use of such bandwidth has been a challenging task in the THz regime, as coupling from the metal-metal waveguides into free space where the wavelength filter is located is difficult because of the geometrical and impedance mismatch between the optical mode and free space. The use of an *LC* circuit would enable an easy coupling to tuning elements, being either inductive or capacitive.

As discussed in the introduction, there is a strong interest in the physics community to study the strong coupling between a cavity and a quantum system. Such studies have already been carried out in the electronic domain at GHz frequencies by using Josephson junctions coupled to superconductive stripline resonators.[14] Clearly, the *LC* circuit is a very interesting object in this context because it does not exhibit any clear size scaling limit other than the thickness of the quantum object it is coupled to. In fact, the area of the capacitor can be continuously scaled down as the thickness is reduced, and the limit to the value of *L* is only the radiative losses. Moreover, the radiative *Q increases* with decreasing circuit size, as shown by two classic antenna studies.[20,21] For quantum optics experiments, Ohmic losses could be quenched by the use of superconductors assuming one operates at a sufficiently low frequency $\nu < 1$ THz.

The spontaneous emission mid-infrared and THz radiation by solid state sources, and in particular by quantum cascade structures, is extremely inefficient because of both the low radiative efficiency of the transition as well as the poor extraction efficiency of high refractive index structures. In an *LC* resonator, the radiative Q can be decreased by a suitable choice of the size of the inductor, and the spontaneous emission can be increased by the Purcell effect. The Purcell factor, measuring the enhancement of the spontaneous emission rate in a cavity with a quality factor Q and a volume V is written as:

$$F_P = \frac{3}{4\pi^2}\frac{Q(\lambda/n)^3}{V}$$

Because the volume of the cavity can be driven orders of magnitudes below λ^3, the Purcell factor could in principle be 10^3–10^4 in a structure designed for this purpose. As a result, efficient room-temperature emission could be achieved as the quantum efficiency could be increased to values close to 10^{-2}–10^{-3}, depending on the wavelength. We believe this approach has many advantages compared to other attempts in this direction that use metallic optical cavities.[22]

We have carried preliminary experiment that demonstrates the feasibility of such a structure. Figure 5 shows a scanning electron micrograph of a possible *LC* resonator consisting of a symmetric arrangement of two capacitors connected by a section of straight wire. The capacitors consist of two metallic planes on each side of a quantum cascade active region designed to operate at the frequency of 1.5–1.6 THz. Electromagnetic finite element simulations of the structure predict a resonant frequency of the structure at 1.52 THz.

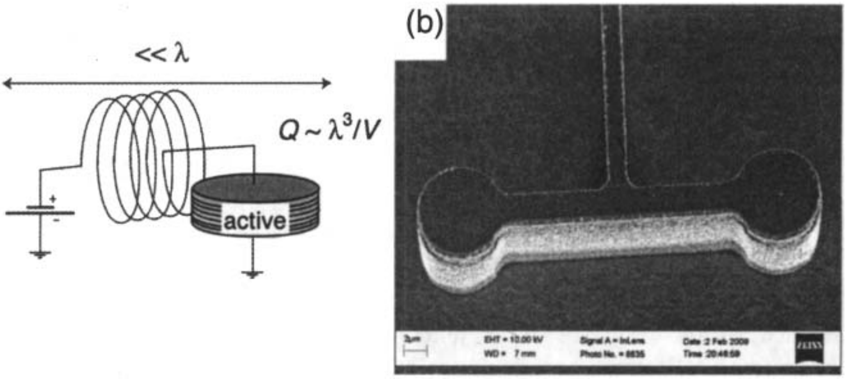


Figure 5. (a) Schematic of diagram of an *LC* resonator; (b) SEM micrograph of an *LC* circuit designed to operate in the THz – overall length of the resonator is 35 μm, much shorter than the wavelength in vacuum (200 μm).

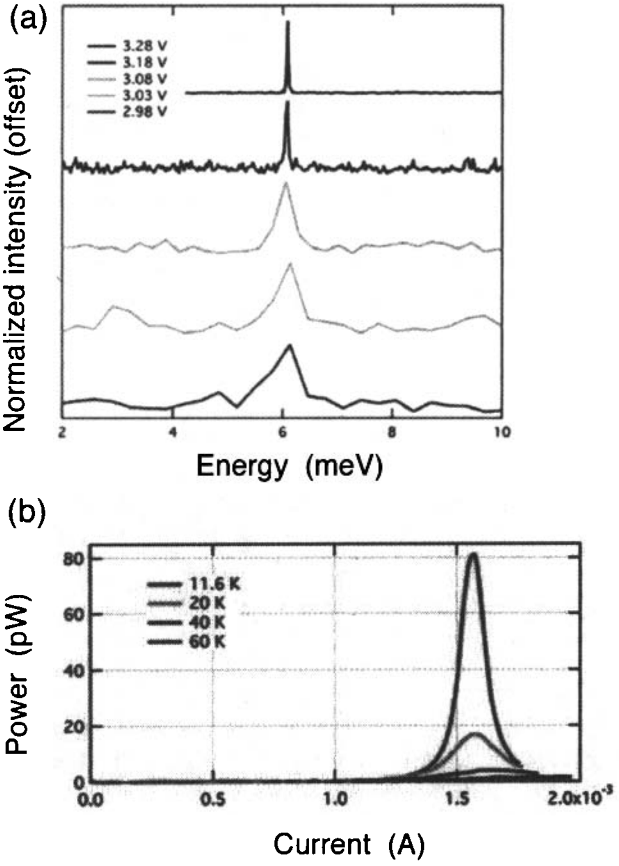


Figure 6. (a) Narrowing of the emission spectrum and (b) light intensity *vs.* current for an *LC* resonator operating at 10 K. Strong oscillation is achieved.

Another structure based on the same concept was operated at cryogenic temperatures and driven by a small continuous current of 1.5 mA. Figure 6 shows the *L–I* curve as well as a spectrum of the device. From the exponential growth of the optical power with current as well as from the narrow emission spectrum, it is clear that self-sustained oscillation has been achieved in the THz by using the coupling between a quantum solid state system and a classical *LC* resonator of deep subwavelength dimension.

5. Conclusions

Since they are based on semiconductor processing techniques, THz QCLs have an excellent potential for the development of new sources in the 1–10 THz frequency range. We have shown that microwave-inspired resonators can lead to devices with very attractive features. A third-order distributed feedback QCL demonstrated single-mode operation, narrow far-field emission, and a high output power. We have also demonstrated the operation of a circuit based on an *LC* circuit where the gain is provided by QCL active material.

Acknowledgments

This work was supported by the Swiss National Foundation under the NCCR Quantum Photonics project.

References

1. P. H. Siegel, "Terahertz technology," *IEEE Trans. Microwave Theory* **50**, 910 (2002).
2. M. Tonouchi, "Cutting-edge terahertz technology," *Nature Photonics* **1**, 97 (2007).
3. P. Haring Bolívar, M. Nagel, F. Richter, M. Brucherseifer, H. Kurz, A. Bosserhoff, and R. Büttner, "Label-free THz sensing of genetic sequences: Towards 'THz biochips'," *Phil. Trans. Royal Soc. A* **362**, 323 (2004).
4. M. Belkin, F. Capasso, F. Xie, A. Belyanin, M. Fischer, A. Wittmann, and J. Faist, "Room temperature terahertz quantum cascade laser source based on intracavity difference-frequency generation," *Appl. Phys. Lett.* **92**, 201101 (2008).
5. J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser," *Science* **264**, 553 (1994).
6. M. Beck, D. Hofstetter, T. Aellen, *et al.*, "Continuous wave operation of a mid-infrared semiconductor laser at room temperature," *Science* **295**, 301 (2002).
7. R. Kohler, A. Tredicucci, F. Beltram, *et al.*, "Terahertz semiconductor-heterostructure laser," *Nature* **417**, 156 (2002).

8. L. Ajili, G. Scalari, D. Hofstetter, *et al.*, "Continuous-wave operation of far-infrared quantum cascade lasers," *Electronics Lett.* **38**, pp. 1675 (2002).
9. G. Scalari, L. Ajili, J. Faist, H. Beere, E. Linfield, D. Ritchie, and G. Davies, "Far-infrared ($\lambda \sim 87$ μm) bound-to-continuum quantum-cascade lasers operating up to 90 K," *Appl. Phys. Lett.* **82**, 3165 (2003).
10. B. S. Williams, S. Kumar, Q. Hu, and J. L. Reno, "Operation of terahertz quantum-cascade lasers at 164 K in pulsed mode and at 117 K in continuous-wave mode," *Opt. Express* **13**, 3331 (2005).
11. S. Kumar, Q. Hu, and J. Reno, "186 K operation of terahertz quantum-cascade lasers based on a diagonal design," *Appl. Phys. Lett.* **94**, 131105 (2009).
12. C. Walther, G. Scalari, J. Faist, H. Beere, and D. Ritchie, "Low frequency terahertz quantum cascade laser operating from 1.6 to 1.8 THz," *Appl. Phys. Lett.* **89**, 231121 (2006).
13. C. Walther, M. Fischer, G. Scalari, R. Terazzi, N. Hoyler, and J. Faist, "Quantum cascade lasers operating from 1.2 to 1.6 THz," *Appl. Phys. Lett.* **91**, 131122 (2007).
14. G. Scalari, N. Hoyler, M. Giovannini, and J. Faist, "Terahertz bound-to-continuum quantum-cascade lasers based on optical-phonon scattering extraction," *Appl. Phys. Lett.* **86**, 181101 (2005).
15. L. Mahler, R. Kohler, A. Tredicucci, *et al.*, "Single-mode operation of terahertz quantum cascade lasers with distributed feedback resonators," *Appl. Phys. Lett.* **84**, 5446 (2004).
16. S. Kumar, B. Williams, Q. Qin, A. W. M. Lee, Q. Hu, and J. Reno, "Surface-emitting distributed feedback terahertz quantum-cascade lasers in metal-metal waveguides," *Opt. Express* **15**, 113 (2007).
17. Y. Chassagneux, R. Colombelli, W. Maineult, *et al.*, "Electrically pumped photonic-crystal terahertz lasers controlled by boundary conditions," *Nature* **457**, 174 (2009).
18. J. A. Fan, M. Belkin, F. Capasso, S. Khanna, M. Lachab, A. G. Davies, and E. H. Linfield, "Surface emitting terahertz quantum cascade laser with a double-metal waveguide," *Opt. Express* **14**, 11672 (2006).
19. K. Roy-Choudhury, S. Haas, and A. F. J. Levi, "Quantum fluctuations in small lasers," *Phys. Rev. Lett.* **102**, 053902 (2009).
20. L. J. Chu, "Physical limitations of small antennas," *J. Appl. Phys.* **19**, 1163 (1948).
21. H. A. Wheeler, "Small antennas," *IEEE Trans. Antenn. Propag.* **23**, 462 (1975).
22. Y. Todorov, I. Sagnes, I. Abram, and C. Minot, "Purcell enhancement of spontaneous emission from quantum cascades inside mirror-grating metal cavities at THz frequencies," *Phys. Rev. Lett.* **99**, 223603 (2007).

Towards Intersubband Polaritonics: How Fast Can Light and Electrons Mate?

A. A. Anappara,[*] **L. Sorba, A. Tredicucci**
NEST, CNR-INFM and Scuola Normale Superiore
Piazza San Silvestro 12, 56127 Pisa, Italy

G. Günter, A. Sell, A. Leitenstorfer, R. Huber
Fachbereich Physik and Center for Applied Photonics, Universität Konstanz
Universitätsstrasse 10, 78464 Konstanz, Germany

S. De Liberato,[†] **C. Ciuti**
Labo. Materieaux et Phénomènes Quantiques, Univ. Paris Diderot-Paris 7
and *CNRS, UMR 7162, Bâtiment Condorcet, 75205 Paris, France*

G. Biasiol
Laboratorio Nazionale TASC CNR-INFM,
Area Science Park, I-34012 Trieste, Italy

1. Introduction

Interaction of light with electronic excitations is a fundamental physical process that can be usually described in terms of absorption or emission of photons. The coupling between a single photonic mode and one elementary transition is quantified by the so-called vacuum Rabi frequency Ω. A strong-coupling regime is achieved when Ω is larger than radiative and nonradiative system loss rates: in this nonperturbative condition, eigenstates are linear superposition of photonic and electronic excitations. This peculiar regime is actively investigated in many research fields, such as ultracold atoms in optical cavities,[1] Cooper-pair boxes in microwave resonators,[2] excitonic transitions in semiconductor microcavities,[3] and surface-plasmon resonators.[4] In the case of atomic transitions, the coupling frequency is much smaller than the transition frequency ω_{12}, which is the other relevant physical scale of the system. On the other hand, in condensed matter, while Ω can be dramatically larger, losses are more significant.[5] Still, it was recently argued that an unprecedented ultrastrong coupling regime can be achieved

---

[*] Present address: Fachbereich Physik and Center for Applied Photonics, Universität Konstanz, Universitätsstrasse 10, 78464 Konstanz, Germany.

[†] Also with Labo. Pierre Aigrain, Ecole Normale Supérieure and CNRS, UMR 8551, 75005 Paris, France.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

when Ω becomes comparable with the transition frequency.[6] Anomalous energy dispersion and a new class of quantum electrodynamics effects were predicted.[7,8] We have investigated quantum-well intersubband excitations in semiconductor microcavities, demonstrating a value of $\Omega_0 = 0.11\omega_{12}$. A peculiar dispersion of the cavity polariton branches was consequently observed,[9] proving that the anti-resonant light-matter coupling and the photon-energy renormalization become directly observable for the vacuum field, in excellent agreement with theory.[6] Furthermore, we have exploited interband ultrafast pumping to optically tune light-matter interaction from weak to ultrastrong and turn on maximum coupling within less than one cycle of light.[10] In this regime, a novel class of extremely nonadiabatic phenomena become observable. In particular, we directly monitor how a coherent photon population converts to cavity polaritons during abrupt switching. This system forms a new and promising laboratory for unprecedented sub-cycle QED effects and represents an efficient room-temperature switching device at the ultimate speed.

2. Intersubband cavity polaritons

Intersubband optical phenomena involve levels originating in a semiconductor heterostructure from the quantum-mechanical confinement of charge carriers in one direction. In *n*-doped quantum wells, intersubband absorption typically occurs in the range from mid-infrared to terahertz (THz) and can be observed with photons of transverse-magnetic (TM) polarization. Energy, carrier density, and matrix elements are the relevant parameters of the resonance, and can be tailored through structural design. Intersubband transitions offer a close similarity to atomic ones, as a consequence of their ideally delta-like joint density of states and of the large separation from continuum levels. The strong coupling with the electromagnetic mode of a planar semiconductor resonator, and the corresponding formation of the so-called *intersubband polaritons*, was observed in GaAs/AlGaAs[5,11,12] and InAs/AlSb[13] material systems, up to room temperature. In an intersubband microcavity of length L_{cav}, embedding N_{QW} quantum wells doped to a carrier density of N_{2DEG}, the coupling frequency can be expressed in Gaussian units as:

$$\Omega_{\text{R},k} = \sqrt{\frac{4\pi\omega_{12}^2 d_{12}^2}{\hbar\varepsilon_{\text{r}} L_{\text{cav}}\,\omega_{\text{cav}}(k)}\,\frac{k^2}{k_{\text{Z}}^2 + k^2}}\,\sqrt{N_{\text{QW}} N_{\text{2DEG}}} \tag{1}$$

where d_{12} is the dipole moment of the transition, ε_{r} is the cavity dielectric constant and k_{Z} is the quantized cavity photon wavevector along the growth direction. At resonance, *i.e.* when the transition frequency ω_{12} is equal to the frequency of the cavity mode $\omega_{\text{cav}}(k)$, the magnitude of the intersubband-polariton splitting $2\hbar\Omega_0 = 2\hbar\Omega_{\text{R},k\text{res}}$ can be maximized by the independent tailoring of these parameters and can become a substantial fraction of the transition energy, thanks also to the low energy and large dipole moment of intersubband excitations. These solid-state systems can be grown by mature epitaxial growth techniques such as molecular beam epitaxy (MBE) and represent optimal candidates to realize the ultrastrong coupling regime of light-matter interaction.

The Hamiltonian of the intersubband microcavity can be written using a bosonic approximation, since the excitation density of the transition (intersubband excitations per unit area of the sample) is very small compared to the density of the two-dimensional electron gas.[6] The Hopfield-like Hamiltonian then takes the form:

$$H = H_{\mathrm{res}} + H_{\mathrm{dia}} + H_{\mathrm{anti\text{-}res}}\,. \tag{2}$$

It consists of three qualitatively different contributions that we shall analyze separately. The first term H_{res} is given by:

$$H_{\mathrm{res}} = \hbar \sum_{k} \left\{ \omega_{\mathrm{cav},k}(a_k^{\dagger} a_k + 1/2) + \omega_{12} b_k^{\dagger} b_k + i\Omega_{\mathrm{R},k}(a_k^{\dagger} b_k - a_k b_k^{\dagger}) \right\} \tag{3}$$

where $a_k^{\dagger}$ (a_k) is the creation (annihilation) operator for the fundamental cavity-photon mode with in-plane wavevector k and frequency $\omega_{\mathrm{cav},k}$ and $b_k^{\dagger}$ (b_k) is the creation (annihilation) operator of the bright intersubband-excitation mode of the doped multiple quantum well structure. H_{res} describes the energy of the bare cavity photon, the intersubband polarization field, and the resonant part of the light-matter interaction (corresponding to the creation/annihilation of one photon with the concomitant annihilation/creation of an intersubband excitation with the same in-plane wavevector).

The middle term in (2) contains the diamagnetic term (proportional to the square of the vector potential A) and gives a renormalization of the photon energy due to the interaction with matter:

$$H_{\mathrm{res}} = \hbar \sum_{k} D_k(a_k^{\dagger} a_k + a_k a_k^{\dagger}) \tag{4}$$

where, for a quantum well, the diamagnetic coupling constant D_k is approximately given by $D_k \approx \Omega_{\mathrm{R},k}^{2}/\omega_{12}$.[6]

The last term in (2) is comprised of the so-called "anti-resonant" terms, corresponding to the simultaneous creation and annihilation of two excitations with opposite in-plane wavevectors:

$$H_{\mathrm{anti\text{-}res}} = \hbar \sum_{k} \left\{ i\Omega_{\mathrm{R},k}(a_k b_{-k} - a_k^{\dagger} b_{-k}^{\dagger}) + D_k(a_k a_{-k} - a_k^{\dagger} a_{-k}^{\dagger}) \right\} \tag{5}$$

Matrix elements of (5) are non-zero only when coupling states with different total number of cavity photons and intersubband excitations. This term is suppressed in first order perturbation theory. Neglecting $H_{\mathrm{anti\text{-}res}}$, the Hamiltonian (2) commutes with the boson number and can be block-diagonalized in a finite dimensional subspace. This kind of approximation is the keystone of all analytical results in the field of light-matter interaction, and is usually known as rotating wave approximation (RWA).[14] Experimental evidence of its violation was found only for the case of dressed states in strongly-driven systems with a large number of photons, by observing energy shifts or forbidden transitions.[14,15]

3. Ultrastrong coupling regime

The elusive H_{dia} and $H_{anti\text{-}res}$ contributions to the interaction with the vacuum field are the hallmark of the ultrastrong-coupling regime and at the origin of the peculiar quantum nature of the states.[6] A simple spectroscopic identification, for instance based on the excitation energies, would be extremely challenging in most microcavity systems. In reality, though, the situation of intersubband microcavities, which use a planar geometry with a resonator designed to operate at oblique incidence, is quite special. Measurements at large angles, in fact, highlight energy deviations of the polariton dispersion, as exemplified in Fig. 1. This effect originates from the fact that different polariton energies calculated for the same k would be observed at different angles in a linear optical experiment.[16]

Polariton energies measured at a fixed angle are those obtained from the intercepts of the $\omega(k)$ polariton curves with the line $\omega = ck/(n\sin\theta)$. For the large experimental angles used, this has a similar slope to a good portion of the polariton dispersion. Small energy deviations between two polariton curves (for instance those calculated with and without the anti-resonant term) produce then larger energy differences in the intercepts, *i.e.* in the peak positions at that given angle.

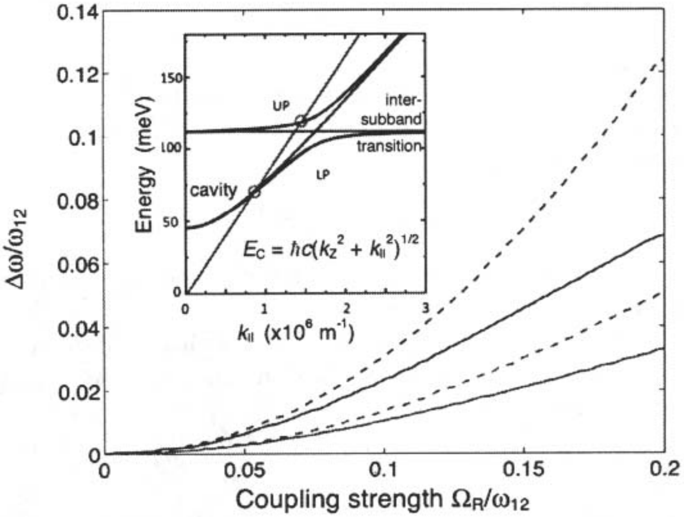


Figure 1. Difference from the full-Hamiltonian eigenvalues of the polariton energies calculated either without the $H_{anti\text{-}res}$ term (lower curves) or without both $H_{anti\text{-}res}$ and H_{dia} (upper curves) plotted as a function of coupling strength (solid and dashed lines refer to the lower and upper polariton branch, respectively). This calculation was performed for a fixed resonant angle of 60°. Clearly, deviations amount to ~5 % already for $\Omega_R/\omega_{12} \sim 0.1$. The inset shows schematically the dispersion of the two polariton branches; the intercepts with the light line at the specific angle correspond to the minima measured in a reflectance experiment.

Optical confinement in the microcavity used in this investigation is based on total internal reflection from a low refractive-index cladding at the bottom and on reflection from a semiconductor-metal interface at the top, see Fig. 2.[17] The heterostructure was grown by solid-source MBE on an undoped GaAs (001) substrate. The cladding region was realized by sandwiching a 1.65 μm AlAs layer between two GaAs layers doped to $5\text{x}10^{18}$ cm^{-3}, each having a thickness of 150 nm. Thickness and doping level of the cladding layers were carefully tuned to ensure a strong optical confinement in the cavity, but at the same time enough coupling with the outside to allow for a clear identification of the polariton features. The active region consists of 70 repeats of *n*-doped 6.5 nm thick GaAs quantum wells separated by 8 nm thick $\text{Al}_{0.35}\text{Ga}_{0.65}\text{As}$ barriers. Layer thicknesses were chosen so as to have only two bound subbands and ensure quantum decoupling of adjacent wells. The doping level ($3.25\text{x}10^{12}$ cm^{-2} in each well) leads to population of the ground state only and to a single intersubband transition.

Light was coupled into the microcavity through the substrate, and the cavity throughput was probed by angle-resolved reflectance measurements. The experimental geometry is detailed in Fig. 2: the sample was mechanically lapped

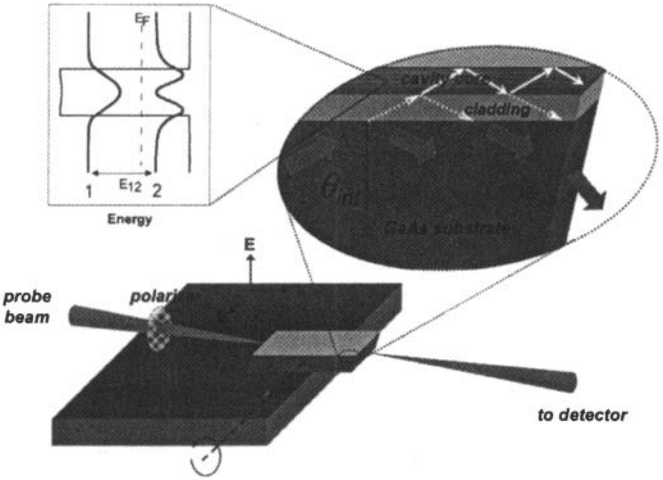


Figure 2. Schematic diagram of the experimental set-up employed for the angle-resolved reflectance measurements. The prism-shaped sample is mounted on a copper block and can be rotated to vary the internal incidence angle. The blown up detail of the waveguide illustrates the working principle of the resonator: large arrows show the optical path of the light in the substrate, dotted white arrows denote evanescent-wave coupling in and out of the resonator, and solid white arrows correspond to intracavity reflections. The cavity core consists of 70 repeats of GaAs/$\text{Al}_{0.35}\text{Ga}_{0.65}\text{As}$ quantum-well units. The band profile and squared moduli of the subband envelope functions of one of the quantum wells are shown in the top left inset.

into a wedge-shaped prism with the polished facets at an angle of 70° with respect to the cavity plane. The prism was soldered onto a copper block and mounted inside the sample compartment of a Nicolet Fourier-transform infrared spectrometer (FTIR), equipped with an Ever-Glo source and a liquid-nitrogen cooled HgCdTe detector. A metallic wire-grid polarizer was inserted in the optical path to select the TM polarization of the probe beam. By manually rotating the sample holder, the angle between the infrared beam and the prism facet could be varied, enabling us to change the incident angle (θ_{int}) on the cavity surface around the central value of 70° defined by the prism shape.

Figure 3(a) shows the angle-resolved TM reflectance spectra in the frequency range of the intersubband transition. The spectra were collected at room temperature, with a resolution of 0.25 meV. Curves display two dips corresponding to the formation of intersubband polaritons. The minimum splitting between the polariton dips is about 90 meV, although a precise anti-crossing point cannot be defined: the two structures shift almost in parallel to higher energies with increasing θ_{int}. This behavior stems again from the fact that polariton peaks at the same internal angle do not correspond to the same *k*.[16] The bare intersubband-transition energy of the active region was measured in another wedge-shaped prism, polished at a 45° angle. The reflectance spectrum was collected at an

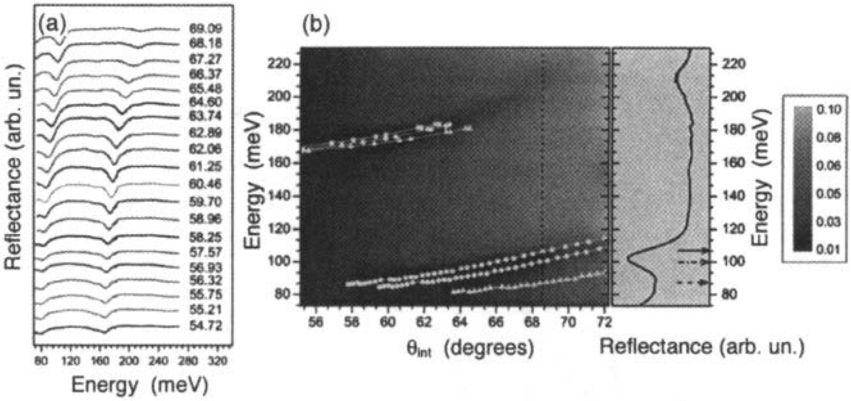


Figure 3. (a) TM reflectance spectra of the microcavity sample wedged at 70° for different internal angles of incidence θ_{int} at the substrate-cavity interface (values given in degrees). Data were collected at 300 K with a resolution of 0.25 meV. Spectra are offset for clarity. (b) Angle-resolved reflectance of the intersubband microcavity displayed as a contour plot in logarithmic scale. The superimposed curves are theoretical calculations with $\hbar\Omega_0$ = 17 meV, computed using the full Hamiltonian (circles), without the anti-resonant term only (squares), and without both photon energy renormalization and the anti-resonant term (triangles). The right panel shows the reflectance spectrum corresponding to the internal angle of 68.5°. Arrows show the calculated lower-polariton energies for the three Hamiltonians detailed above.

internal angle of about 37°, which excluded any cavity-induced shift of the intersubband absorption. The recorded transition energy is 152 meV, with a full width at half maximum (FWHM) of about 12 meV.

Since the bottom mirror utilizes total internal reflection, one cannot determine precisely the cavity resonance energy through measurements at zero incidence, where the intersubband transition does not couple to the radiation. In order to avoid unnecessary fitting parameters or relying on the uncertain knowledge of dielectric constants and growth rates, we decided to use a second sample, identical in the growth sequence, but without any doping in the active region. Angle-resolved reflectance measurements conducted in this latter sample were then used to determine the energy dispersion of the cavity mode. The quality of the growth and the thickness difference between the two samples were checked using X-ray diffraction (XRD). The reference sample was also wedged at an angle of 70°; Scanning electron microscope (SEM) images of the cleaved facets of the two wedge-shaped samples were recorded to check the angular difference between the mechanically polished facets. Having measured both the intersubband transition energy and the angle-dependent cavity mode frequency, we proceeded to model the data, with the vacuum Rabi frequency as the only free fitting parameter. In order to prove that our intersubband microcavity is indeed in the ultrastrong coupling regime, we compare in Fig. 3(b) the measured reflectance with the computed polariton dispersion curves. The energies calculated from the full Hamiltonian (2) are shown as white circles. A perfect agreement is found for a vacuum Rabi energy $\hbar\Omega_0 = 17$ meV, ~11 % of the intersubband transition energy.[9] Note that, as discussed in Ref. 16, the actual value of the vacuum Rabi energy is much smaller than half the splitting observed in the spectra. The triangles in Fig. 3(b) were instead obtained neglecting both H_{dia} and $H_{\text{anti-res}}$ terms, while the squares exclude only the anti-resonant contribution but include the photon renormalization. Though only small deviations appear for the upper polariton, the lower polariton dispersion calculated with either of the latter approximations is clearly incompatible with the experimental measurements. In order to better quantify this difference, the 68.5° spectrum is analyzed in detail in the right panel of Fig. 3(b). The energy derived without $H_{\text{anti-res}}$ is 6 meV above the measured polariton dip, while that without H_{dia} and $H_{\text{anti-res}}$ is 13 meV below, both well beyond the experimental resolution.

4. Nonadiabatic switching

In the ultrastrong coupling regime, the anti-resonant terms of the interaction become relevant already in equilibrium; as a consequence, the ground state of the system is a squeezed vacuum containing a finite number of virtual photons.[6] As long as the vacuum is confined inside the cavity, its quantum nature may be regarded as an academic question. Theory, however, shows that nonadiabatic switching of Ω_R may release these virtual quanta in correlated pairs.[6,8] As a matter of fact, nonadiabatic control poses an extreme experimental challenge: changes are required to be induced on a scale set by the cycle of light. Electronic tuning via an external gate voltage[17,18] is clearly too slow; theoretical proposals to modulate the dielectric constant of a cavity or the reflectivity of a mirror have not been implemented for the required time scale either.

To overcome this challenge, we implemented an all-optical scheme for femtosecond control of ultrastrongly coupled cavity polaritons – a first practical test-bed to target nonadiabatic QED phenomena.[10] The concept is illustrated in Fig. 4. For maximum modulation, we start with an empty microcavity and turn on light-matter coupling by direct injection of the emitter itself. Our sample contains 50 identical, undoped GaAs QWs separated by $Al_{0.33}Ga_{0.67}As$ barriers. Radiative transitions between the subbands only become possible if electrons are injected extrinsically. The intersubband absorption line is centered at $\hbar\omega_{12}$ = 113 meV (wavelength λ = 11 μm). The effective thickness of the structure corresponds to $\lambda/2$ at an internal propagation angle of $\theta = 65°$. We employ a 12-fs control pulse centered at a photon energy of 1.55 eV[19] to excite electrons resonantly from the valence band into conduction subband $|1\rangle$ (vertical beam in Fig. 4 and thick arrow in the inset), activating the intersubband field within femtoseconds.

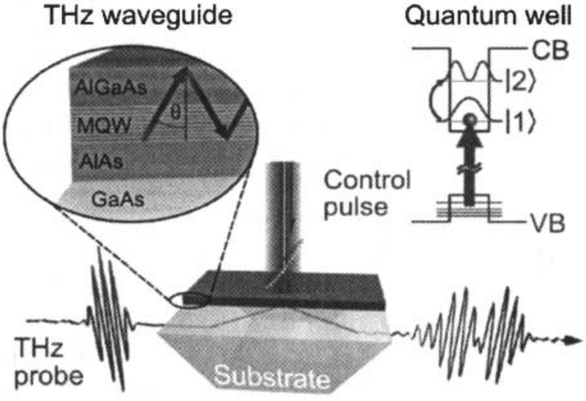


Figure 4. Femtosecond switch of ultrastrong light-matter coupling. A bare microcavity features minimal reflectance $R(\omega)$ at the photon resonance ω_C. After introduction of a degenerate quantum emitter, vacuum fluctuations of the photon field are coherently absorbed and reemitted at a rate Ω_R, giving rise to anticrossing cavity polaritons. The 50 undoped GaAs quantum wells are embedded into a planar waveguide structure based on total internal reflection on either side. The QWs are positioned at the field antinode. Electronic transitions between subbands $|1\rangle$ and $|2\rangle$ are activated by near infrared 12-fs control pulses (vertical beam) populating level $|1\rangle$. Intersubband transitions may then resonantly couple to TM polarized mid-infrared cavity photons propagating at $\theta = 65°$. Few-cycle multi-THz transients guided through the prism-shaped substrate are reflected off the waveguide to probe the ultrafast buildup of the mixed light-matter system. The pulse front of the near-infrared control is tilted (broken circle in control beam) to match the geometry of the phase planes of the probe.

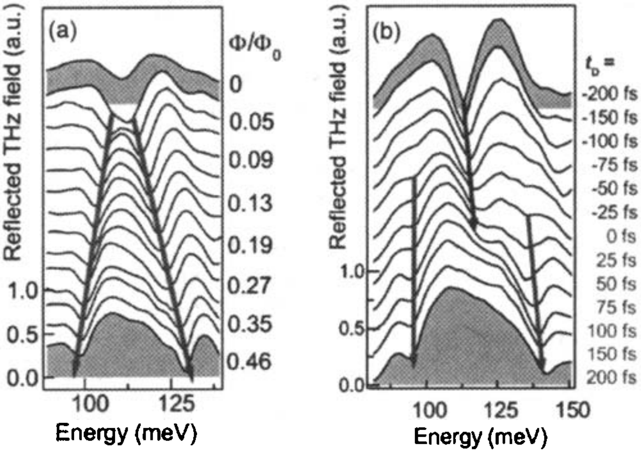


Figure 5. (a) THz reflectance spectra measured at room temperature for various fluences Φ (vertically offset) of the control pulse (t_D = 20 ps). Minima indicate eigenmodes of the system. (b) Spectra of the reflected THz field are given for various delay times t_D (offset vertically). The 12-fs control pulse ($\Phi_0 = 0.1$ mJ/cm^2) arrives at $t_D = 0$. Central arrow: bare cavity resonance, lateral arrows: ultrastrongly coupled intersubband cavity polariton branches.

The eigenmodes of this system are probed by resonant scattering of a second light pulse: a TM polarized phase-locked mid-infrared transient (oscillating curve in Fig. 4) is coupled through the prism-shaped substrate and reflected off the photoexcited area of the waveguide. In order to achieve sub-cycle temporal resolution, we map out the real-time oscillations of the field of the reflected pulse by phase-matched electro-optic detection.[20] Eigenmodes of the cavity are identified via characteristic minima in the Fourier spectra of the reflected transients. All experiments are performed under ambient conditions.

Figure 5(a) demonstrates that Ω_R is continuously tunable via the control fluence Φ: a series of spectra is recorded for a fixed delay time t_D = 20 ps between the near-infrared control and the multi-THz probe pulse. In equilibrium ($\Phi = 0$), a single reflectance minimum at $\hbar\omega_C$ = 113 meV (top curve) attests to the only resonance of the unexcited cavity, the bare photon mode. With increasing fluence, the system traverses all three regimes of light-matter interaction: starting with a broadening of the cavity line in the weak coupling regime ($\Phi \leq 0.03\Phi_0$, where Φ_0 = 0.1 mJ/cm^2), Ω_R exceeds the widths (FWHM ~ 5 meV) of intersubband and cavity resonances already for $\Phi > 0.07\Phi_0$, and two strongly coupled cavity polariton branches become clearly discernible. Further increase of the fluence enhances the separation of the minima to 50 meV, corresponding to 44% of the bare photon frequency. This value is comparable to what has been achieved before in doped structures and fulfils the criteria of ultrastrong coupling. We expect the scheme to be further scalable by using a higher control fluence.

The central question however is: how rapidly may ultrastrong coupling be activated? Figure 5(b) displays amplitude spectra recorded at various delay times t_D for $\Phi = \Phi_0$. For $t_D \leq -50$ fs the cavity resonance (center arrow) is clearly seen with a minimum amplitude reflectivity below 10%. The control pulse induces dramatic spectral changes of order unity, on the femtosecond scale. The initial bare photon state is replaced by two coupled polariton branches appearing simultaneously at energy positions of 93 meV and 143 meV (side arrows in Fig. 5(b)). Most remarkably, the new resonances do not gradually develop out of the bare cavity mode. The system does not sequentially pass through all three regimes of light-matter coupling – a scenario that would manifest itself in a bifurcation similar to Fig. 5(a). Instead, switching from the bare resonance to ultrastrongly coupled polaritons occurs discontinuously. It appears that cavity-intersubband coupling forms quasi-instantaneously once level $|1\rangle$ is populated. Surprisingly, the bare photon mode suffers a distinct broadening already at negative delay times $t_D \geq -100$ fs. Moreover, in the neighborhood of $t_D = 0$ fs, broadened versions of bare photon and polariton resonances seem to be present simultaneously.

These paradoxical features are fingerprints of a novel nonadiabatic phenomenon which we like to term "perturbed cavity emission decay". The physics is most striking in the time domain, illustrated in Fig. 6.[10] When a few-cycle multi-THz transient – the time trace in Fig. 6(a) – impinges on the unexcited

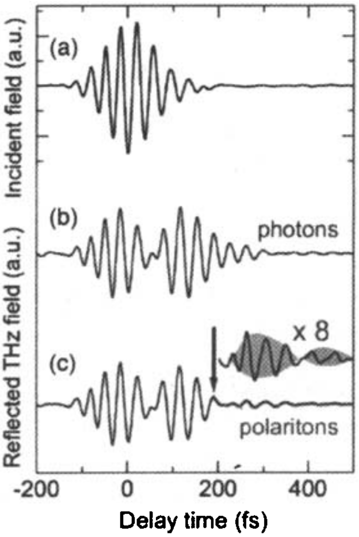


Figure 6. (a) Few-cycle transient incident on the microcavity structure; (b) time trace of the reflected THz field *vs.* time *t* (c) a 12-fs control pulse arrives within the coherence window of the cavity mode (vertical arrow). The reflected THz field traces the nonadiabatic switch-on of ultrastrong coupling: the control pulse abruptly changes the exponential emission decay into a characteristic beating signature.

modulator structure, part of its energy is directly reflected off the cavity surface. A second portion is transmitted into the structure, prepares a coherent photon state in the resonator, and gets reemitted subsequently, contributing to the overall reflected signal. This dynamics is encoded in a characteristic twin-pulse temporal structure of the reflected transient – see Fig. 6(b). The most intriguing situation is realized if we switch the eigenstates of the cavity while a coherent state of bare photons is still present, as in Fig. 6(c): the incident THz transient first prepares a photon population in the unexcited cavity. During the free emission decay (second burst of the reflected field), the control pulse abruptly alters the eigenstates of the system. Remarkably, the decay of the bare cavity mode is interrupted on a time scale shorter than the oscillation cycle of light – a compelling proof of nonadiabaticity. The subsequently reemitted transient exhibits a characteristic two-mode beating in the time domain. The corresponding spectrum displays minima at photon energies of 93 meV and 143 meV – the hallmarks of both polariton branches. The initial coherent photon population is thus effectively converted into ultrastrongly coupled cavity polaritons beating at the splitting frequency. The fact that both branches of the new eigenstates are populated simultaneously confirms that switching indeed proceeds nonadiabatically.

5. Conclusions

The experiments described in this chapter open up a new domain of light-matter interaction at the ultrastrong and ultrafast limit. We have provided evidence that anti-resonant light-matter coupling and photon-energy renormalization can become significant even in the interaction with the vacuum electromagnetic field of a microcavity. Furthermore we have observed the nonadiabatic dynamics of a coherent photon population during sub-cycle switching. Such experiments provide a benchmark for latest cavity-QED theories in the ultrastrong coupling regime and encourage a systematic quest for nonadiabatic quantum optical phenomena.

References

1. Y. Colombe, T. Steinmetz, G. Dubois, F. Linke, D. Hunger, and J. Reichel, "Strong atom-field coupling for Bose-Einstein condensates in an optical cavity on a chip," *Nature* **450**, 272 (2007).
2. A. Wallraff, D. I. Schuster, A. Blais, *et al.*, "Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics," *Nature* **431**, 162 (2004).
3. K. Hennessy, A. Badolato, M. Winger, *et al.*, "Quantum nature of a strongly coupled single quantum dot–cavity system," *Nature* **445**, 896 (2007).
4. J. Bellessa, C. Bonnand, J. C. Plenet, and J. Mugneir, "Strong coupling between surface plasmons and excitons in an organic semiconductor," *Phys. Rev. Lett.* **93**, 036404 (2004).

5. D. Dini, R. Köhler, A. Tredicucci, G. Biasiol, and L. Sorba, "Microcavity polariton splitting of intersubband transitions," *Phys. Rev. Lett* **90**, 116401 (2003).
6. C. Ciuti, G. Bastard, and I. Carusotto, "Quantum vacuum properties of the intersubband cavity polariton field," *Phys. Rev. B* **72**, 115303 (2005).
7. C. Ciuti and I. Carusotto, "Input-output theory of cavities in the ultrastrong coupling regime: The case of time-independent cavity parameters," *Phys. Rev. A* **74**, 033811 (2006).
8. S. De Liberato, C. Ciuti, and I. Carusotto, "Quantum vacuum radiation spectra from a semiconductor microcavity with a time-modulated vacuum Rabi frequency," *Phys. Rev. Lett.* **98**, 103602 (2007).
9. A. Anappara, S. De Liberato, A. Tredicucci, *et al.*, "Signatures of the ultrastrong light–matter coupling regime," *Phys. Rev. B* **79**, 201303 (2009).
10. G. Günter, A. A. Anappara, J. Hees, *et al.*, "Sub-cycle switch-on of ultrastrong light–matter interaction," *Nature* **458**, 178 (2009).
11. E. Dupont, J. A. Gupta, and H. C. Liu, "Giant vacuum-field Rabi splitting of intersubband transitions in multiple quantum wells," *Phys. Rev. B* **75**, 205325 (2007).
12 L. Sapienza, A. Vasanelli, C. Ciuti, C. Manquest, C. Sirtori, R. Colombelli, and U. Gennser, "Photovoltaic probe of cavity polaritons in a quantum cascade structure," *Appl. Phys. Lett.* **90**, 201101 (2007).
13. A. A. Anappara, D. Barate, A. Tredicucci, J. Devenson, R. Teissier, and A. Baranov, "Giant intersubband polariton splitting in InAs/AlSb microcavities," *Solid State Commun.* **142**, 311 (2007).
14. E. K. Irish, "Generalized rotating-wave approximation for arbitrarily large coupling," *Phys. Rev. Lett.* ***99*** 173601 (2007), and references therein.
15. S. Hofferberth, B. Fischer, T. Schumm, J. Schmiedmayer, and I. Lesanovsky, "Ultracold atoms in radio-frequency dressed potentials beyond the rotating-wave approximation," *Phys. Rev. A* **76**, 013401 (2007).
16. A. A. Anappara, A. Tredicucci, F. Beltram, *et al.*, "Cavity polaritons from excited subband transitions," *Appl. Phys. Lett.* **91**, 231118 (2007).
17. A. A. Anappara, A. Tredicucci, G. Biasiol, and L. Sorba, "Electrical control of polariton coupling in intersubband microcavities," *Appl. Phys. Lett.* **87**, 051105 (2005).
18. A. A. Anappara, A. Tredicucci, F. Beltram, G. Biasiol, and L. Sorba, "Tunnel-assisted manipulation of intersubband polaritons in asymmetric coupled quantum wells," *Appl. Phys. Lett.* **89**, 171109 (2006).
19. R. Huber, F. Adler, A. Leitenstorfer, M. Beutter, P. Baum, and E. Riedle, "12 fs pulses from a cw-pumped 200 nJ Ti:sapphire amplifier at a variable repetition rate of up to 4 MHz," *Opt. Lett.* **28**, 2118 (2003).
20. C. Kübler, R. Huber, S. Tübel, and A. Leitenstorfer, "Ultrabroadband detection of multi-THz field transients with GaSe electro-optic sensors: Approaching the near infrared," *Appl. Phys. Lett.* **85**, 3360 (2004).

Si_3N_4/SiO_2 Planar Photonic Structures Fabricated by Focused Ion Beam

L. A. M. Barea, F. Vallini, D. L. S. Figueira, A. Da Silva Filho, N. C. Frateschi
Device Research Laboratory, Applied Physics Dept. Gleb Wataghin,
Univ. of Campinas–UNICAMP, 13083-970 Campinas, SP, Brazil

A. R. Vaz
Center for Semiconductor Components,
University of Campinas–UNICAMP, 13083-870 Campinas, SP, Brazil

1. Introduction

Micro-optical resonators are essential components in optical integrated systems, as they facilitate the selective capture, release, and storage of photons as well as interaction with matter and other photons. As an example, microdisks[1] and micro-rings[2,3] are elements that provide extremely long photon lifetimes that can be used for spectral filtering and for highly efficient nonlinear photon interactions.[4,5] The small volumes and the high photonic lifetime in micro-resonators are very important, for example, to enable four-wave mixing (FWM) with a single pumping frequency.[6]

However, coupling light to and from the micro-resonator is still achieved via hybrid integration, typically using tapered silica fibers.[7] In this chapter, we present a new technique for the fabrication of structures with monolithically integrated resonators and planar tapered waveguides, all based on SiN_x/SiO_2 layers deposited on silicon.

We use an approach that combines nanofabrication techniques based on focused ion beam (FIB) milling and smoothing[8] and conventional microfabrication techniques. We have fabricated resonators coupled to tapered waveguides using this approach. Tapered waveguides with 0.8 dB/cm losses were obtained. Finally, we observed the photon generation at a micro-ring resonance due to FWM generated by pumping at a subsequent micro-ring resonance.

2. Fabrication and characterization

All structures were fabricated on a silicon (100) substrate. First, a 0.5 μm SiO_2 layer is obtained by wet oxidation.[9] Subsequently, we deposited 0.3 μm of silicon nitride (Si_3N_4) over the SiO_2 layer using an electron cyclotron resonance (ECR) plasma deposition (process parameters: 125 sccm of SiH_4, 20 sccm of Ar and 2.5 sccm of N_2 at 3 mtorr process pressure with 250 W of rf power for 40 min).

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

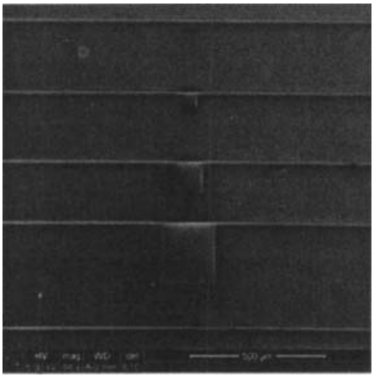


Figure 1. SEM micrograph of Si_3N_4 structures fabricated by conventional microfabrication techniques. The square mesas will be used for the fabrication of resonators.

Fabrication of tapered waveguides coupled to the resonators was accomplished using a hybrid technique involving conventional fabrication and FIB milling. The first step consisted of fabricating Si_3N_4 ridges connected to square mesas. The ridges were 10 μm wide, whereas the mesas ranged from 25 to 200 μm on the side. Figure 1 shows an SEM micrograph of the fabricated structures. The Si_3N_4 was etched chemically in a buffered oxide etchant for 8 minutes.

The second step of the fabrication involved FIB milling of the squares to form resonators and of the waveguide structures near the squares to form adiabatic tapering for field delocalization. The FIB was performed using a dual-beam FIB/SEM 200 model from FEI, with gallium as the ion source. The inset in Fig. 2(a) shows the programmed milling pattern. The milling is done with an emission current of 50 pA and a voltage of 30 kV. The low current provides high resolution (20 nm) for the process. Figure 2(a) shows a SEM micrograph of a microdisk resonator (50 μm radius) coupled to a tapered waveguide (0.5 μm wide). The tapered 0.2 μm wide waist is aligned with the resonator center. The taper length is 300 μm. The time to complete the milling was 45 min. Figure 2(b) shows a microring resonator obtained with the milling of the internal region of the microdisk described above. The ring has 0.5 μm width and the gap between the resonator and the guide is 0.2 μm.

The last step of the fabrication consisted of a 2 μm SiO_2 layer deposition over the entire structure. This step was done using the ECR system with 200 sccm of SiH_4, 20sccm of O_2, and 20 sccm of Ar at 10 mtorr; 750 W of rf power for 10 min. The structures are then cleaved orthogonally to the waveguide direction.

The expected guiding and coupling behavior of the structures can be predicted by using a 2D finite element method (FEM) of the Comsol Multiphysics™ package. Figure 3(a) shows a schematic drawing of the Si_3N_4 waveguide structure used in the simulation, whereas Fig. 3(b) shows the simulated transverse electric (TE) modes. The waveguide is multimode, as expected. Figure 3(c) shows the

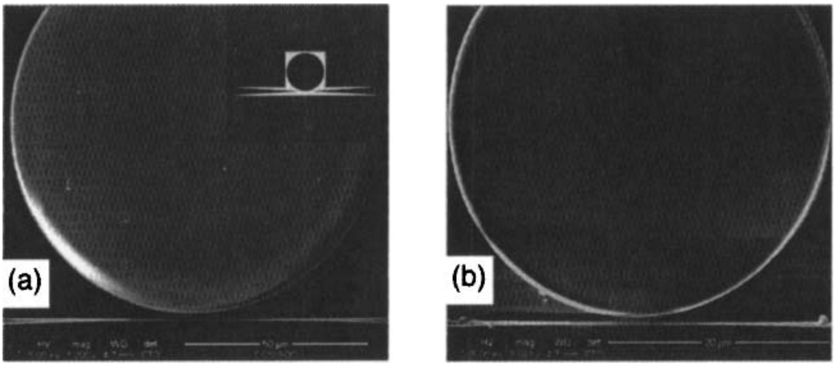


Figure 2. (a) SEM micrograph of a 50 µm radius microdisk 0.2 µm distant from the 0.5 µm wide tapered waveguide waist (inset shows the ion milling pattern); (b) SEM micrograph of a 0.5 µm width micro-ring (50 µm radius) fabricated on the same disk.

light confinement in the tapered region of the waveguide. The wide evanescence field allows coupling to structures as far as 1 µm from the waist. Obviously, a narrower tapered region gives a higher evanescence field, as shown in Fig. 3(d) for a 0.2 µm waist. Still, in order to best match the phase velocity between the waveguide and the ring we fabricated both structures with the same width of 0.5 µm.

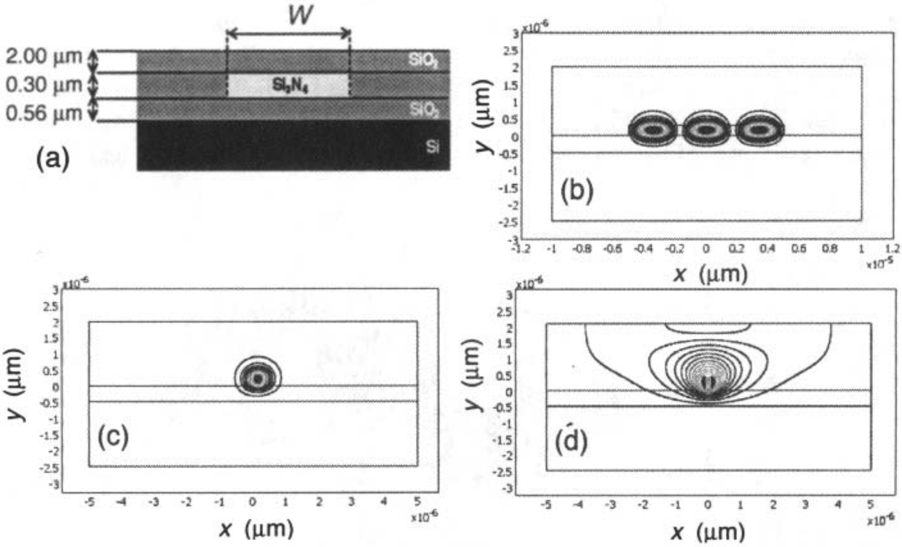


Figure 3. (a) Schematic drawing of the Si_3N_4 waveguide structure; (b–d) electric field contours for the TE mode for waveguides with width $W = 10$ µm (b), 0.5 µm (c), and 0.2 µm (d).

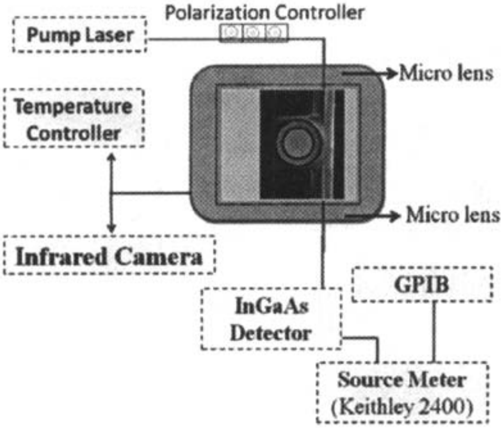


Figure 4. Experimental set-up.

Figure 4 shows the experimental set-up used for the optical characterization of our structures. The sample was positioned over a block with a length smaller than the waveguide length (about 1 mm). This block was set over a Peltier element for temperature control. A multi-channel tunable laser (1500 to 1600 nm) and a polarization controller were used as the light source. A GRIN rod lens for collimation and a micro lens for focus were used to couple light to and from the planar structures and the fiber. The signal output was sent to an InGaAs detector.

Before measuring our resonators, we fabricated a 2500 μm long waveguide with a 300 μm long tapered region of 0.2 μm width to evaluate the transmission losses. Figure 5 shows the spectrum of this waveguide. Fabry-Perot oscillations with a free spectral range (FSR) of 0.28 nm are observed. Losses of 0.6 cm^{-1} are obtained using the Hakki-Paoli method,[10] assuming a reflectivity of 12% at the facets based on the Fresnel coefficient. This corresponds to 0.8 dB/cm loss.

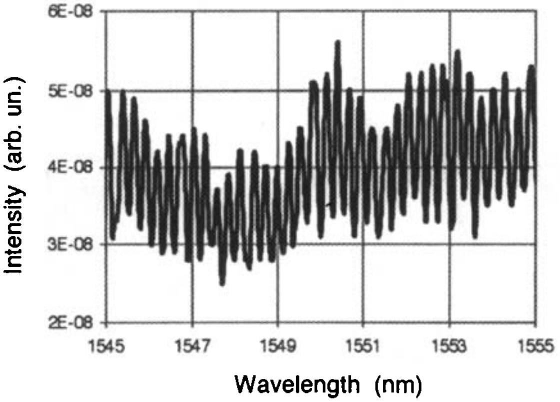


Figure 5. Spectrum of a 2500 μm long and 10 μm wide waveguide with a 300 μm tapered region and a waist of 0.2 μm. The associated FRS is 0.28 nm.

Transmission measurements for the waveguide coupled to resonator structures were subsequently performed on a 2100 μm long waveguide coupled to a 25 μm radius micro-ring with the dimensions described above. The solid line in Fig. 6(a) shows the transmission spectrum. We observe two transmission minima at 1551.0 nm and 1558.6 nm. The dotted line shows the simulated transmission spectrum.

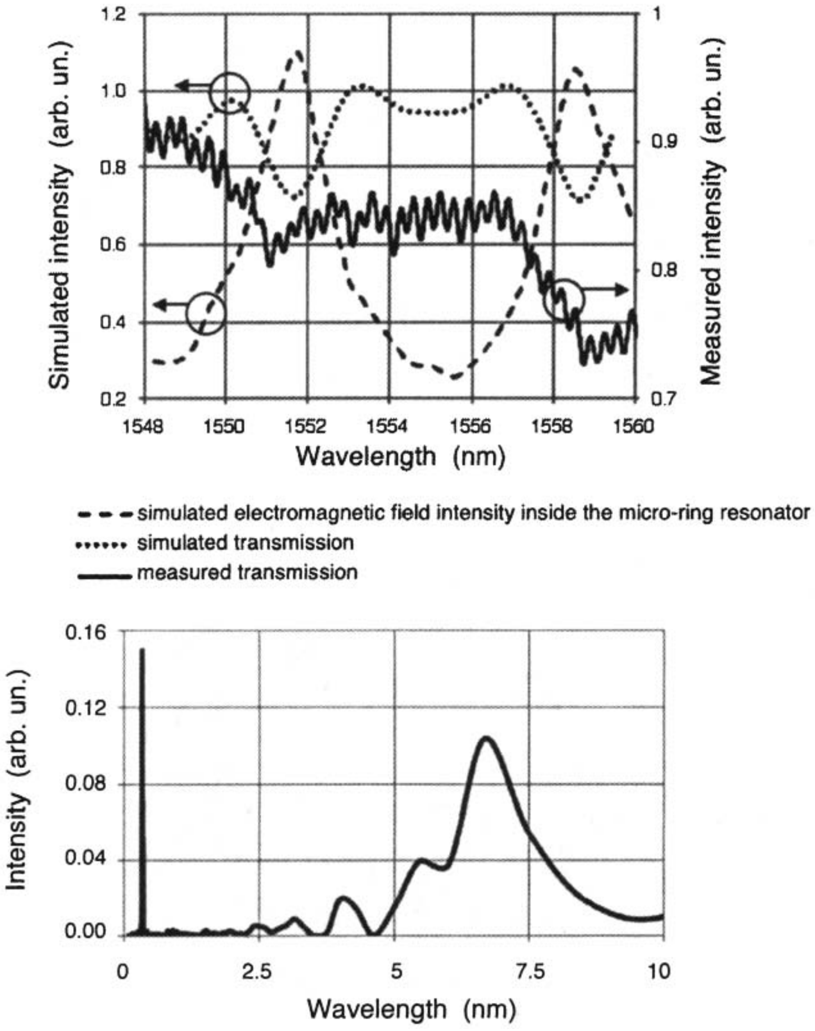


Figure 6. Transmission spectrum of a micro-ring resonator with 25 μm radius and 0.5 μm width coupled to a tapered waveguide of 0.5 μm waist at a distance of 0.2 μm (top); Fourier transform of the spectrum (bottom).

The dashed line shows the simulated electromagnetic field intensity inside the micro-ring resonator. The experimentally observed minima match the expected micro-ring resonances. As expected, the transmission minima occur at the micro-ring maxima, in good agreement with the simulation. Figure 6(b) shows the Fourier transform of the transmission spectrum. There are two maxima: one at 0.3 nm related to the FSR of the Fabry-Perot cavity and another at 7.6 nm related with the FSR of the micro-ring resonator coupled to the tapered waveguide. This result is in agreement with simulations of the 25 μm radius disk.

In order to study photon creation within the resonator, we chose one of the resonances for pumping the micro-ring resonator and repeated the transmission experiment. Figure 7 shows the transmission spectrum (dotted line) obtained previously and a new spectrum for a 1558.6 nm pump signal with constant power of 2 dBm and a varying probe signal from 1545 to 1560 nm with constant power of 6 dBm (solid line). The enhanced transmission at 1551 nm is a result of the generation of extra photons at this resonance wavelength. The generation of extra photons at this resonance is most probably due to the pumping at the subsequent micro-ring mode at 1558.6 nm. This is possible since four-wave mixing between neighboring modes in micro-ring resonators is allowed for materials with small chromatic dispersion. In these cases, both the angular momentum and the photon energy vary linearly with the azimuthal quantum number.[5,11]

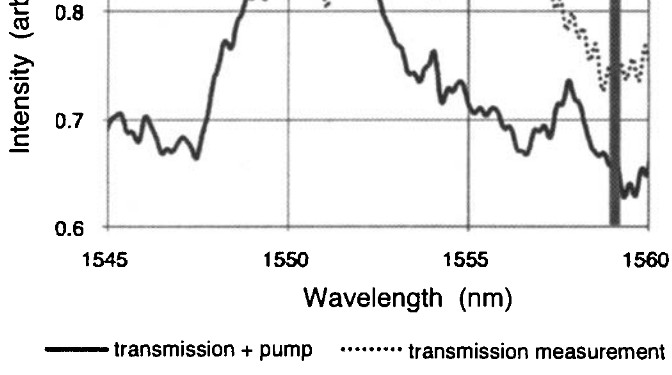


Figure 7. Transmission spectra: dotted line shows the transmission without pumping; solid line shows the transmission with 2 dBm pumping at the 1558.6 nm resonance.

5. Conclusions

We have presented a new technique for the fabrication of Si_3N_4/SiO_2 planar resonators coupled to tapered waveguides. This new technique consists of conventional microfabrication combined with high-resolution focused ion beam processing. Tapered waveguides with typical loss of 0.8 dB/cm were obtained. Also, we observed photon generation at a micro-ring resonance due to four-wave mixing generated by pumping at a subsequent micro-ring resonance.

Acknowledgments

The authors would like to acknowledge the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), the Centro de Pesquisa em Óptica e Fotônica (CEPOF) and Institutos Nacionais de Ciência e Tecnologia (INCT) with Fotônica para Comunicações Ópticas (FOTONICOM), for the financial support. The authors would like to thank José Godoy Filho for assistance in the oxidation process.

References

1. P. E. Barclay, K. Srinivasan, O. Painter, B. Lev, and H. Mabuchi, "Integration of fiber-coupled high-Q SiN_x microdisks with atom chips," *Appl. Phys. Lett.* **89**, 131108 (2006).
2. T. Barwicz, M. Popovic, P. Rakich, M. Watts, H. Haus, E. Ippen, and H. Smith, "Micro-ring-resonator-based add-drop filters in SiN: fabrication and analysis," *Opt. Express* **12**, 1437 (2004).
3. S. Zheng, H. Chen, and A. Poon, "Micro-ring-resonator cross-connect filters in silicon nitride: Rib waveguide dimensions dependence," *IEEE J. Select. Topics Quantum Electronics* **12**, 1380 (2006).
4. K. J. Vahala. "Optical microcavities," *Nature* **424**, 839 (2003).
5. P. Del'Haye, A. Schliesser, O. Arcizet, T. Wilken, R. Holzwarth, and T. J. Kippenberg, "Optical frequency comb generation from a monolithic microresonator," *Nature* **450**, 1214 (2007).
6. K. Inoue, "Four-wave mixing in an optical fiber in the zero-dispersion wavelength region," *J. Lightwave Technol.* **10**, 1553 (1992).
7. S. M. Spillane, T. J. Kippenberg, O. J. Painter, and K. J. Vahala, "Ideality in a fiber-taper-coupled microresonator system for application to cavity quantum electrodynamics," *Phys. Rev. Lett.* **91**, 043902 (2003).
8. F. Vallini, D. S. L. Figueira, P. F. Jarschel, L. A. M. Barea, A. A. G. Von Zuben, and N. C. Frateschi, "Effects of Ga^+ milling on InGaAsP quantum well laser with mirrors milled by focused ion beam", *J. Vac. Sci. Technol. B* **27**, L25 (2009).
9. K. K. Lee, D. R. Lim, L. C. Kimerling, J. Shin, and F. Cerrina, "Fabrication of ultralow-loss Si/SiO_2 waveguides by roughness reduction," *Optics Lett.* **26**, 1888 (2001).
10. B. W. Hakki and T. L. Paoli, "Gain spectra in GaAs double-heterostructure injection lasers," *J. Appl. Phys.* **46**, 1299 (1975).
11. M. K. Chin and S. T. Ho, "Design and modeling of waveguide-coupled single-mode micro-ring resonators," *J. Lightwave Technol.* **16**, 1433 (1998).

Part II

Electronic Devices and Systems

This page intentionally left blank

2 Electronic Devices and Systems

Although silicon is undoubtedly still full of steam, the word "post-CMOS" has become a commonplace. With the end of the CMOS-scaling evolution in sight, no one at the FTM seemed to be in the panic mode. Perhaps, instead of seeing the end of a profession based on near-perfection in precision engineering, the FTM researchers see great opportunities opening up in front. Indeed, a number of fascinating concepts were discussed, from Mott transistors to hybrid superconductor-CMOS memory. Limits of the current memory technologies and prospects of alternative memory mechanisms were analyzed. Alternative engineering and system strategies, such as token-passing architectures and electron mass engineering, were examined, often from refreshing perspectives such as energy efficiency and info-carrying distance, *etc.*

As often is the case, new technologies are enabled by discoveries and advances in materials. The big-brother silicon is clearly pressed (at least in terms of the conference publicity) by its much nimbler sibling carbon. Germanium is another column-IV element that is challenging silicon at its own game, while away from column-IV materials, chalcogenide glassy semiconductors were proposed for memory applications.

Contributors

2.1 F. Balestra

2.2 P. M. Solomon

2.3 W. Van Den Daele, S. Cristoloveanu, E. Augendre, C. Le Royer, J.-F. Damlencourt, D. Kazazis, and A. Zaslavsky

2.4 Enrico Sangiorgi, Marco Braccioli, and Claudio Fiegna

2.5 D. Grützmacher, Th. Schäpers, S. Mantl, S. Feste, Q. T. Zhao, H. Hardtdegen, R. Calarco, M. Lepsa, and N. Demarina

2.6 T. Baron, C. Agraffeil, F. Dhalluin, M. Kogelschtaz, G. Cunge, T. Chevolleau, B. Salem, B. Salhi, H. Abed, A. Potié, L. Latu-Romain, C. Ternon, K. Aissou, L. Montès, Mur, G. Molas, B. De Salvo, E. Jalaguier, T. Ernst, P. Ferret, P. Gentile, and N. Pauc

2.7 N. V. Demarina and D. Grützmacher

2.8 Qiliang Li, X. Zhu, Y. Yang, D. E. Ioannou, J. S. Suehle, and C. A. Richter

2.9 Jang-Sik Lee

2.10 K. D. Tsendin

2.11 M. Dorojevets

2.12 M. E. Levinshtein and S. L. Rumyantsev

This page intentionally left blank

Silicon-Based Devices and Materials for Nanoscale CMOS and Beyond-CMOS

F. Balestra
Sinano Institute, IMEP-Minatec, CNRS-Grenoble INP
3 Parvis Louis Néel, BP 257, 38016 Grenoble, France

1. Introduction

The minimum critical feature size of the elementary devices (physical gate length L_G of the transistors) will drop from 25 nm in 2007 (65 nm technology node) to 4.5 nm in 2022 (11 nm technology node). In the sub-10nm range, beyond-CMOS devices will certainly play an important role and could be integrated on CMOS platforms in order to pursue integration down to nm structures. While Si will remain the main semiconductor material for the foreseeable future, the needed performance improvements for the end of the ITRS Roadmap[1,2] will substantially expand the number of materials, technologies and device architectures.

Silicon-on-insulator (SOI)-based devices seem to be the best candidates for the ultimate integration of integrated circuits on silicon.[3,4] An overview of the main advantages of SOI for the nanoelectronics of the next two or three decades is presented in this chapter. We will consider nanoscale CMOS, emerging and beyond-CMOS nanodevices, based on innovative concepts, technologies and device architectures. The flexibility of the SOI structure enables new device architectures that can be used to optimize electrical properties for both low-power and high-performance circuits. These transistors are also promising for high frequency and memory applications. We will discuss the performance and physical mechanisms in single- and multi-gate/multi-channel thin film MOSFETs. A comparison between the performance of Si, Ge and III-V semiconductors on insulator will be given. The impact of tensile or compressive strains in the channel, of high-κ materials and metal gates as well as metallic Schottky source-drain architectures will be discussed. The promise of beyond-CMOS nanodevices on SOI, based on nanowires or small slope switch structures will also be presented. Finally, the advantages of using multi-gate architectures for scaling down DRAM, SRAM and non-volatile memories will be outlined.

2. More Moore

- *New channel materials for ultimate CMOS*

As simple scaling of silicon CMOS becomes increasingly complex and expensive, there is considerable interest in increasing performance by using

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

strained channels, which reduces the valley degeneracy and lowers the carrier effective mass, thus improving carrier mobility and drive current in a device. The combination of the advantages of SOI structures for improving the electrostatic control and of strained semiconductors is very promising. An original method for strained SOI (SSOI) is based on thin SiGe buffer layers relaxed by He^+ ion implantation and post-implantation annealing. If the Si cap layer thickness lies below a critical thickness, an induced strain is transferred immediately during the relaxation of the SiGe layer. The wafer is then cleaved with hydrogen implantation according to the Smart-Cut® process and the stack layer is bonded to an oxidized handle wafer. Split-CV (capacitance-voltage) mobility measurements have shown enhanced electron mobility values as high as 1200 $cm^2/V{\cdot}s$.[5]

Using this biaxial tensile stress as starting material, uniaxial tensile strain can be obtained by lateral strain relaxation of patterned structures – see Fig. 1 (a), (b). Figure 1(c) shows the transfer characteristics of a strained and an unstrained nanowire (NW) *n*-FET. The NWs have a square cross-section of 40x40 nm and a length of 3 μm. The subthreshold slope of both devices is between 70–80mV/dec. Both devices have a very low I_{OFF} and a high I_{ON}/I_{OFF} ratio. The on-current of the uniaxially strained device is enhanced on the average by a factor of 2.5 due to uniaxial tensile strain, while strain does not degrade the low off-current. The inset of Fig. 1(c) shows plots $I_D/(g_M)^{1/2}$ *vs.* V_G for a strained and unstrained device measured at $V_D = 50$ mV. The slope of the linear part of the curves is related to the carrier mobility. The mobility enhancement due to uniaxial tensile strain is found to be ~2.3. For maximum performance enhancement due to lateral strain relaxation SSOI devices must have a large length to width ratio.[6]

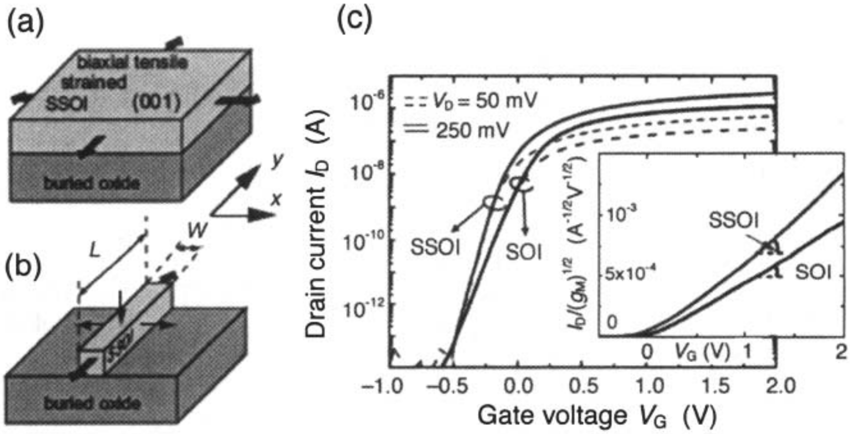


Figure 1. (a) 50nm SSOI with 1GPa biaxial tensile stress is used as starting material. (b) Uniaxial tensile strain is obtained by lateral strain relaxation of patterned structures. (c) Transfer characteristics of two NW-FETs, one fabricated on SOI and one on uniaxial SSOI. The channel length of both devices is L_G = 3 μm and the gate oxide thickness is 5nm. The inset shows the $I_D/(g_M)^{1/2}$ plot for the devices. The slope of the linear region is related to carrier mobility.[6]

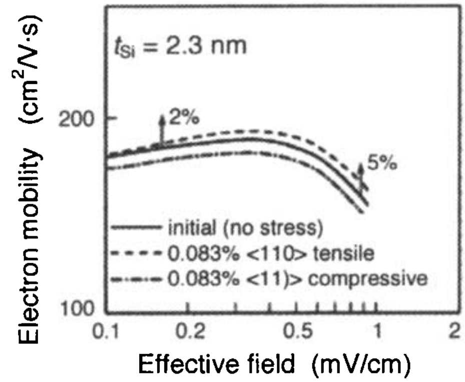


Figure 2. Electron mobility in 2.3-nm ultra-thin-body MOSFET under <110> uniaxial strain.[7]

Strain engineering is also useful for mobility enhancement for Si film thickness in the sub-10nm range,[7] which will be needed for the ultimate integration of nanoscale MOSFETs. A similar enhancement of electron mobility in 3.5 nm SOI devices under biaxial and uniaxial tensile strain has been obtained. The electron mobility is also enhanced in 2.3 nm Si layer under uniaxial tensile strain, as shown in Fig. 2, whereas hole mobility increases in a 2.5 nm Si film under uniaxial compressive strain.

The impact of the strain for ultra-short and ultra-thin fully-depleted (FD) SOI is shown in Fig. 3. A 40% gain in I_{ON} for a given I_{OFF} for L_G = 18 nm devices is obtained for various t_{Si}. The mobility increases down to ultra-thin t_{Si} and up to high channel carrier density. The ballisticity in SOI and SSOI devices is similar. The improvement of the injection velocity (about 35%) in SSOI nMOS, holds down to t_{Si} ~ 2.5 nm, explains the substantial I_{ON} increase shown in Fig. 3.[8]

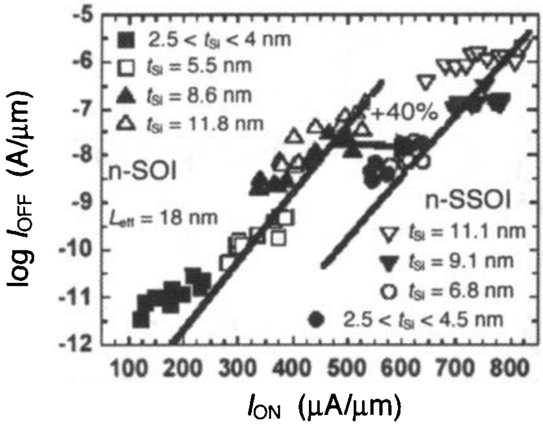


Figure 3. I_{ON} *vs.* I_{OFF} showing the impact of tensile strain obtained via the Smart-Cut® process from strained Si/relaxed $Si_{0.8}Ge_{0.2}$.[8]

- *Metallic Schottky source/drain MOSFETs*

As CMOS technology is entering the decananometer era, the contact resistance associated with the silicide/silicon interface is identified as one of the biggest challenges to solve in order to preserve current drive capabilities. In that context, source/drain engineering takes an increasing importance in the development of leading edge CMOS generations, because of the increasing impact of S/D series resistances on transistor performance. In order to further pursue downscaling of MOSFETs in the $L_G < 32$ nm range, one alternative is to implement metallic S/D contacts combined to a dopant segregation strategy at a reduced thermal budget. The expected benefit is to considerably reduce the specific contact resistance of the metal/semiconductor junction while keeping activated dopants sharply localized at the interface. The efficiency of dopant-segregated Schottky contacts has been demonstrated.[9,10] It has been shown that a sub-100 meV barrier can be obtained consistently with the boron pile-up observed at the PtSi/Si interface. A new state-of-the-art current drive performance has been established for Schottky barrier MOSFETs at $L_G = 25$ nm: $I_{ON} = 530$ µA/µm at $V_G = V_D = -1.1$ V. Figure 4 demonstrates that metallic S/D is competitive with the best unstrained SOI pMOS technologies. A record rf performance for a 30 nm *p*-type unstrained thin-film FD-SOI Schottky barrier MOSFET has been demonstrated with an f_T of 180 GHz.[9]

- *High-κ materials*

In the search for new insulator materials for the 16 nm CMOS node and beyond, the dielectric constant itself is not sufficient as a physical quantity to meet the technical device specification. Even more important is the product C_E of the

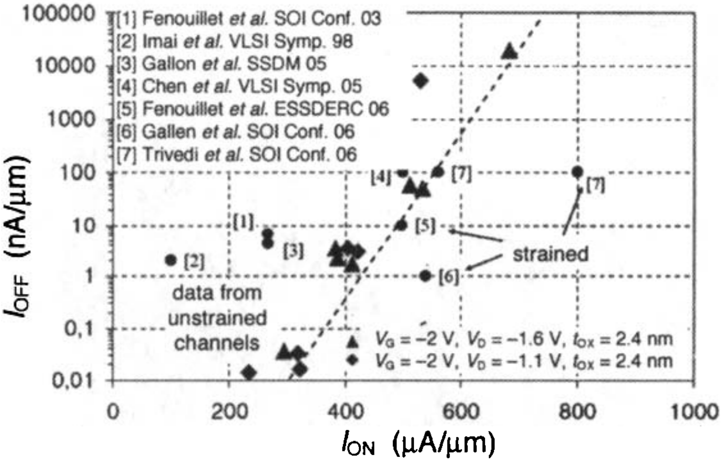


Figure 4. I_{ON} *vs.* I_{OFF} in state-of-the art of S/D pMOS on SOI indicating that boron dopant-segregated devices lead both Schottky barrier and conventional unstrained thin-film SOI technologies.[9]

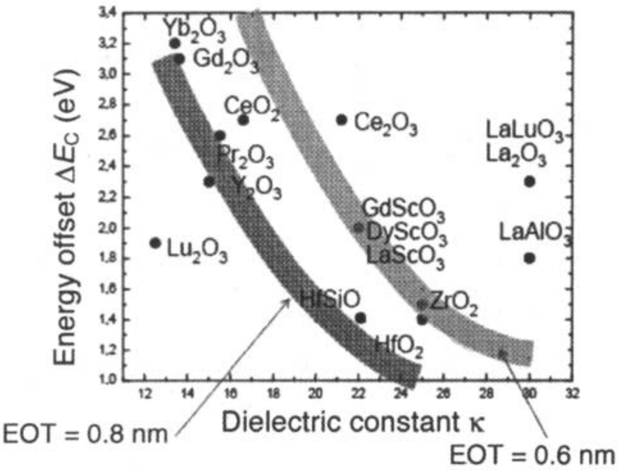


Figure 5. Energy offset values between the conduction bands of various dielectrics and silicon *vs.* dielectric constant. The upper area is a border for the 16 nm high-performance DG CMOS SOI node. The lower area represents the corresponding border for 16 nm low operating power DG CMOS SOI technology.[11]

dielectric constant κ and the energy barrier height governed by the offset ΔE between the energy bands of the insulator and the silicon, $C_E = \kappa\Delta E$. For bulk CMOS nodes, C_E is larger than that needed for thin double-gate SOI technology. In Fig. 5 these relations are summarized for high-performance and low operating power DG SOI MOSFETs. A large number of possible gate dielectrics can be used for multi-gate devices. Thermal stability and charge carrier trapping will govern the choice of these future gate dielectrics.

- *Multi-gate devices*

Multi-gate MOSFETs realized on thin films are the most promising devices for the ultimate integration of MOS structures due to the volume inversion or volume accumulation in the conductive layer (for enhancement- and depletion-type devices, respectively),[12] leading to an increase of the carrier density and mobility, and hence higher current drive, optimum subthreshold swing, and better control of short-channel effects and I_{OFF}. The latter is particularly significant because I_{OFF} is the main challenge for future nanodevices due to static power consumption.

Figure 6 shows a comparison of short channel effects for bulk, single-gate SOI, single-gate silicon-on-nothing (SON), and double-gate (DG) MOSFETs. For $L_G <$ 30 nm integration, the advantages of SOI and multi-gate architectures are clear.[13]

The I_{ON} of the MOSFET is limited to a maximum value I_{BL} that is reached in the ballistic transport regime. Figure 7 plots the self-consistent Monte Carlo simulation of the ballistic ratio BR = I_{ON}/I_{BL} against the drain-induced barrier lowering (DIBL), which is one of the main short channel effects. Evidently, one can increase the BR by scaling L_G and thus increasing the longitudinal field at the source, but this comes at the cost of a larger DIBL. For a given DIBL, an increased ballisticity is obtained for low-doped DG SOI devices.[14]

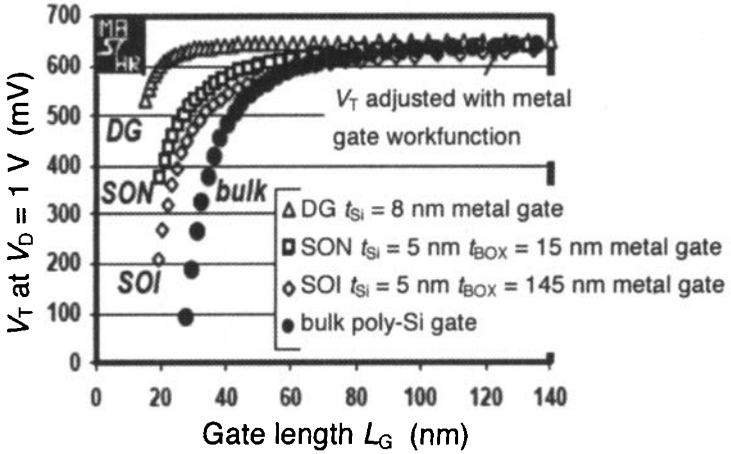


Figure 6. Comparison of threshold voltage V_T *vs.* L_G for various MOSFET architectures from bulk to multi-gate devices.[13]

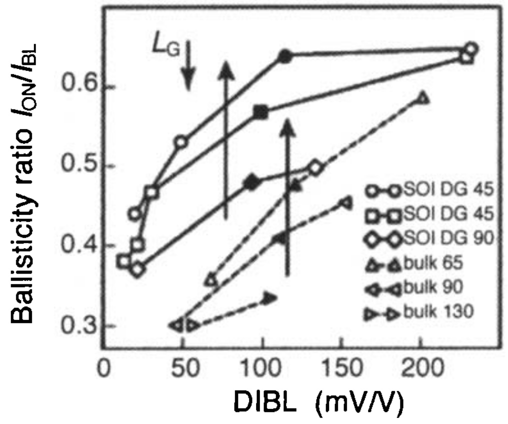


Figure 7. Ballisticity ratio at $V_G = V_D = V_{DD}$ *vs.* DIBL. Filled symbols represent transistors with the nominal L_G for the high-performance MOSFET of each technology node.[14]

The transfer characteristics of several multiple-gate (1–4 gate) MOSFETs, calculated using the 3D Schrödinger-Poisson equation and the nonequilibrium Green's function (NEGF) formalism for the ballistic transport or Monte Carlo simulations, have shown similar trends. The best performance (drain current, subthreshold swing) is obtained in four-gate or gate all around (GAA) structure, see Fig. 8(a).[15,16] However, as Fig. 8(b) demonstrates, the propagation delay in triple-gate and quadruple-gate MOSFETs is degraded due to a strong rise of the gate capacitance. A properly designed DG structure appears to be the best compromise at a given I_{OFF}.[16]

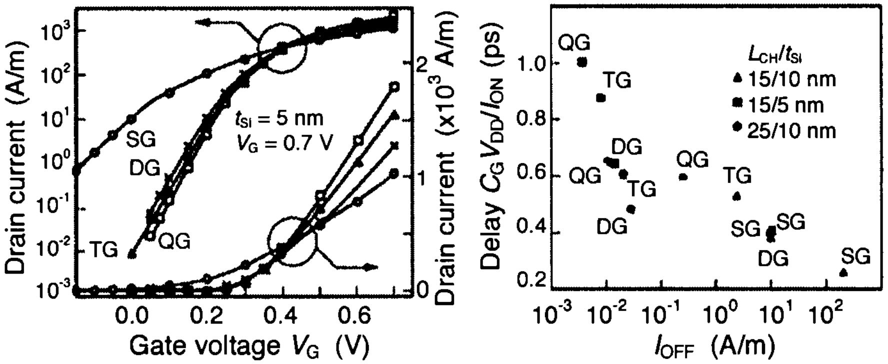


Figure 8. (a) $I_D(V_G)$ at V_D = 0.7 V in thin layers for different multi-gate architectures; (b) propagation delay *vs.* I_{OFF} for single-gate and multi-gate SOI devices.[16]

In decananometer MOSFETs, gate underlap is also a promising solution in order to reduce the DIBL effect. Figure 9 presents the variations of the current drive I_{ON}, the subthreshold current I_{SUB} and the gate direct tunneling current I_{GDT} as a function of gate underlap.[17] The I_{ON} is almost unaffected by the gate underlap, whereas the leakage currents are substantially reduced due to a decrease in DIBL and I_{GDT}. A reduction of the effective gate capacitance C_G for larger underlap values at the same I_{ON} has also been observed. This reduction of C_G leads to a decrease in the propagation delay and power consumption.

Another interesting issue is the comparison of channel materials. Figure 10 plots I_{ON} as a function of L_G for Si, strained Si, Ge and GaAs *n*-channel DG MOSFETs, for a given I_{OFF} for the next 3 generations of ITRS high-performance ICs. The four materials are chosen in their optimized orientation and short channel effects and access resistance are included. When source-drain tunneling is

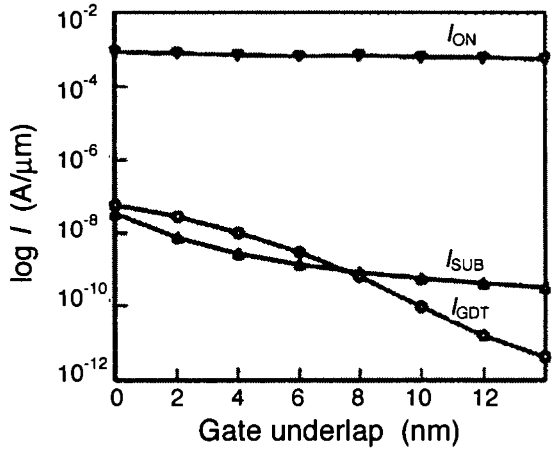


Figure 9. I_{ON}, subthreshold I_{SUB} and gate direct tunneling I_{GDT} *vs.* gate underlap.

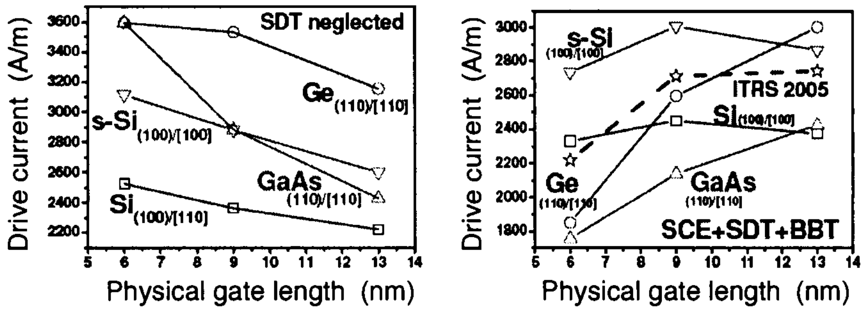


Figure 10. Simulated I_{ON} of n-channel DG MOSFETs *vs.* L_G, including for short channel effects, for different channel materials (Si, strained Si, Ge, GaAs) without (a) and with (b) source-drain tunneling.[18]

neglected, Ge and GaAs devices promise the highest I_D for L_G < 10nm. However, when source-drain tunneling is included, only DG strained Si MOSFETs satisfy the need of ITRS in the sub-decananometer range.[18]

In order to reach very high performance at the end of the roadmap, multi-channel FETs (MCFETs) shown in Fig. 11 present very high current drive, larger than that of GAA devices and exceeding the ITRS requirements. Reported performance of 50 nm NMOS with TiN/HfO$_2$ gate stack reached I_{ON} = 2.27 mA/µm, I_{OFF} = 16.4 pA/µm, and S = 63 mV/decade.[19,20]

- *Variability*

Silicon-on-insulator devices can also be interesting for reducing the variability in decananometer MOSFETs, which also represents a major challenge at the end of the roadmap. The sources of local and inter-die threshold voltage V_T variability in undoped ultra-thin FD-SOI MOSFETs with a high-κ/metal stack have been

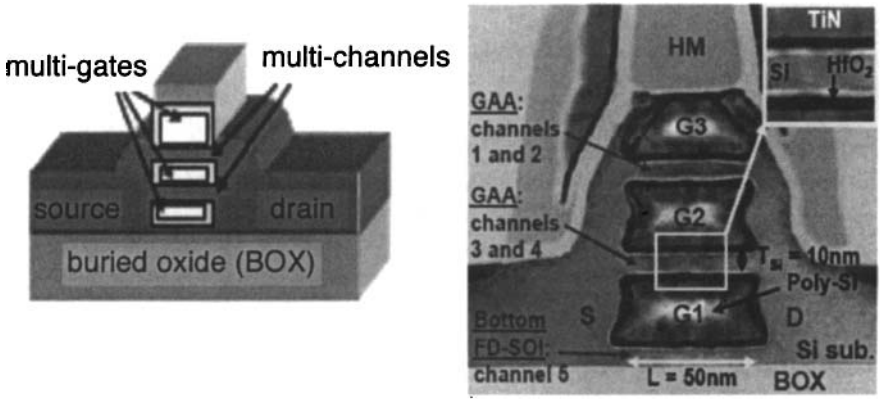


Figure 11. Multi-channels realized using SON technology leading to very high drain currents with excellent control of leakage currents.[20]

experimentally investigated and discriminated. Charges in the gate dielectric and/or TiN gate workfunction fluctuations have been found to be major contributors to the local V_T variability, whereas SOI thickness variations have a negligible impact down to $t_{Si} = 7$ nm. Moreover, t_{Si} scaling has been shown to limit both local and inter-die V_T variability induced by L_G fluctuations. The highest matching performance ever reported for $L_G = 25$ nm MOSFETs has been achieved, demonstrating the effectiveness of the undoped ultra-thin FD-SOI architecture in terms of V_T variability control.[21]

High immunity to threshold variability due to undoped ultra-thin SOI has also been demonstrated in multi-gate SOI devices, which are also more tolerant to line edge roughness-induced variability.[22]

3. Emerging and beyond-CMOS nanodevices

The objectives in this domain are to explore the horizon beyond conventional CMOS – sometimes called beyond Moore – in order to overcome possible downscaling or performance limits in the next two decades. In this field, SOI structures can also be considered as a platform for pursuing integration in the nanoworld.

- *Nanowires*

It has been shown previously that multi-gate architectures based on the concept of volume inversion are very promising in order to overcome the number of CMOS performance challenges down to at least decananometer devices. Gate-all-around semiconductor nanowires (NWs) can be seen as the ultimate integration of these innovative nanodevices and present very interesting properties down to the sub-10 nm range.

The combination of strain effects with NWs can lead to very high performance ICs for the end of the roadmap. Top-down bent GAA nanowires have been fabricated in order to improve carrier mobility and current drive.[23] The NW bending is induced by thermal oxidation in suspended NWs. A maximum tensile strain is obtained in the middle of the wire, raging from 200 MPa to 2 GPa. A substantial enhancement of electron mobility in these structures has been demonstrated.

Nanowire 3D integration is a promising solution to boost the current drive while maintaining a low I_{OFF}. Using the SON process, three stacked GAA sub-15 nm nanowires have been fabricated – see Fig. 12. These NWs were 100 nm long, with 1.8 nm EOT gate insulator (high κ/metal gate stack). They exhibited extremely high current drive and low leakage current: $I_{ON} = 6.5$ mA/μm, $I_{OFF} = 27$ nA/μm for nMOS; $I_{ON} = 3.3$ mA/μm, $I_{OFF} = 0.5$ nA/μm for pMOS. Figure 12 also illustrates a new independent gate fin-FET-like NW structure named ΦFET that has recently been reported, promising new design flexibility. These 3D structures can be extended to a combined vertical and lateral integration for logic, memories and NEMS applications – see Fig. 12. Compared to a standard fin-FET processed alongside, these 3D-NWFET and ΦFETs relax the channel width requirement for a targeted DIBL and exhibit better transport properties (Fig. 13). The ΦFET also exhibits significant performance boosts compared to independent-gate fin-FETs: a two-decade smaller I_{OFF} and a lower subthreshold slope (82 mV/decade instead of 95 mV/decade). This highlights the better scalability of 3D-NWFET and ΦFET compared to standard and insulated-gate fin-FETs.[24,25]

- *Small slope switches*

Even though the aggressive scaling will continue to play an important role in the future nanoelectronics, new technology drivers and new functionalities will open alternative ways for future high performance systems.

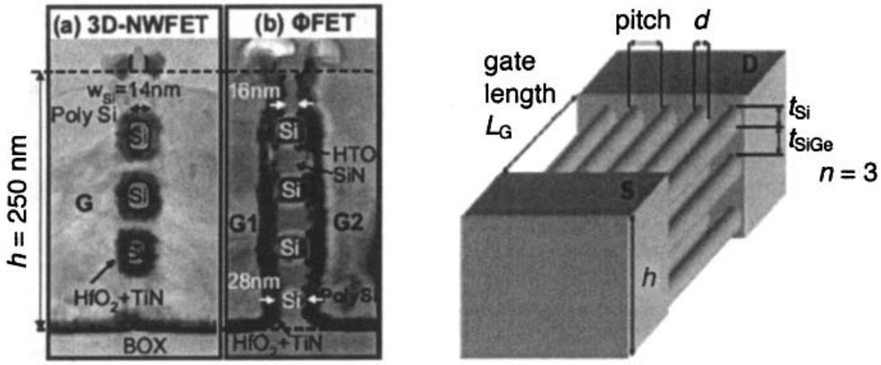


Figure 12. Cross-sectional pictures of a 3D-NWFET (a) and a ΦFET (b) (spacers are introduced in the ΦFET to obtain stacked NWs with independent gate operation); (c) lateral and vertical integration of 3D NWs.[24,25]

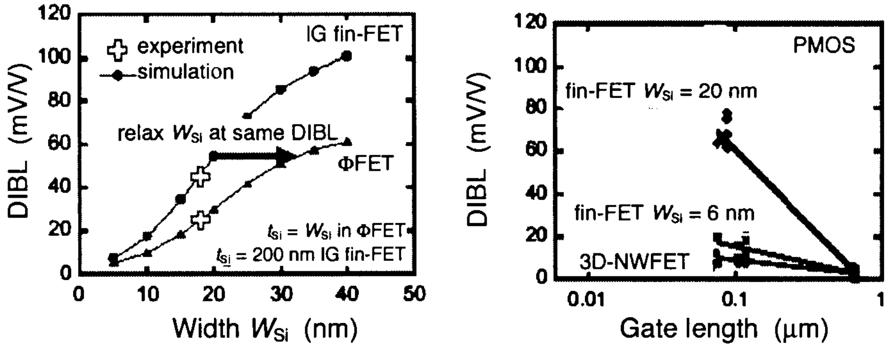


Figure 13. (a) Simulated and measured DIBL *vs.* Si width W_{Si} for ΦFET and insulated-gate fin-FET, L_G = 50 nm; (b) DIBL *vs.* L_G for 3D-NWFET compared to a standard fin-FET, showing strong short-channel effect reduction in NWFETs.[24]

One interesting class of beyond-CMOS devices consists of the small slope switches. A small slope electronic switch is defined as a solid-state semiconductor device showing a value of the subthreshold slope S smaller than the 60mV/decade limit for a conventional MOSFET, set by the Boltzmann distribution at room temperature. The smaller the value of S, the more abrupt the transition from the off-state to the on-state. Benefits of small slope switches are the ultra-low standby power due to a very low I_{OFF} but also the high-speed potential and dynamic power savings, since less power is drawn per transition when the subthreshold slope is more abrupt. Development of small slope switches requires exploration of new physical principles for very abrupt off-on transition, such as: (i) impact ionization; (ii) band-to-band tunneling; (iii) electro-mechanical instability; or iv) ferroelectric gate electrodes. It is worth noting that all these small slope device architectures can be implemented as extensions of advanced silicon CMOS or by hybridization of silicon CMOS with other compatible technologies (SiGe, nanowire, CNT, nano-electro-mechanical structures, *etc.*).

The tunnel FET (TFET)[26-28] is a gated *p–i–n* diode with a gate over the intrinsic region; it exploits the gate-controlled electron tunneling from the valence band of the *p*-region to the conduction band of the *i*-region for reverse bias, resulting in very abrupt off-on transition. They have been reported on Si, III-V and CNT alternatives. Multi-gate TFETs have recently shown to present very interesting performance. Double-gate high-κ TFETs can be scaled to shorter lengths before important characteristics such as transconductance, $I_{\mathrm{ON}}/I_{\mathrm{OFF}}$, and subthreshold swing are degraded – see Fig. 14. Optimized silicon body of 7-8 nm (for L_G = 50 nm) in DG TFETs with high-κ dielectrics leads to improved characteristics with higher I_{ON} ($I_{\mathrm{ON}}/I_{\mathrm{OFF}} \sim 10^{11}$) and smaller S (average S = 57 mV/decade; minimum point S = 11 mV/decade).[28]

Double-gate TFETs with strained-Ge heterostructure channel have recently shown high current drive and good S lower than 60 mV/decade, due to small bandgap and tunnel resistance of strained Ge and DG electrostatics.

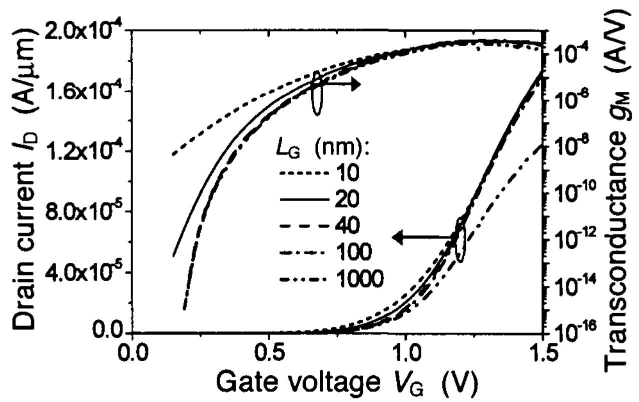


Figure 14. Effect of gate length scaling from 1 μm down to 10 nm on simulated $I_D(V_G)$ and $g_M(V_G)$ characteristics of high-κ DG TFETs.[28]

Using a TFET simulator, the scalability of symmetric and asymmetric DG TFET configurations, shown in Fig. 15, has been studied in order to achieve high I_{ON}/I_{OFF} ratios and remove the ambipolarity of the transfer characteristics that plagues symmetric TFET structures. Simulations show that increasing the drain underlap and lowering the drain doping can kill the ambipolar behavior. These approaches are not fundamentally different and both lower leakage due to a long depletion width and lower electric field on the drain side. Unfortunately, both these TFET structures need a very long non-scaling drain of > 30 nm to achieve a $S < 60$ mV/decade. This puts a severe restriction on device scalability. The ambipolar problem can also be solved using a lateral heterostructure of a large E_G material (like Si) at the drain side, see Fig. 15(d). This drain-side heterostructure appears to be the most scalable approach.[29]

A novel transistor design, based on a fin-FET architecture, which utilizes positive feedback to achieve steep switching behavior has also been proposed – see Fig. 16(a) and (b). The feedback FET exhibits very low subthreshold swing (~2 mV/decade) and high I_{ON}/I_{OFF} ratio (~10^8) to allow for significant reductions in gate voltage swing to < 0.5 V. It is a new candidate to replace the MOSFET for future low power electronic devices and can operate as an *n*- or *p*-channel device. Steep switching behavior is achieved by dynamically reducing V_T as the transistor is turned on, via barrier-height modulation by carrier accumulation in the drain- and source-offset regions inducing a positive feedback and an abrupt switch into the on-state, as shown in Fig. 16(c) and (d).[30]

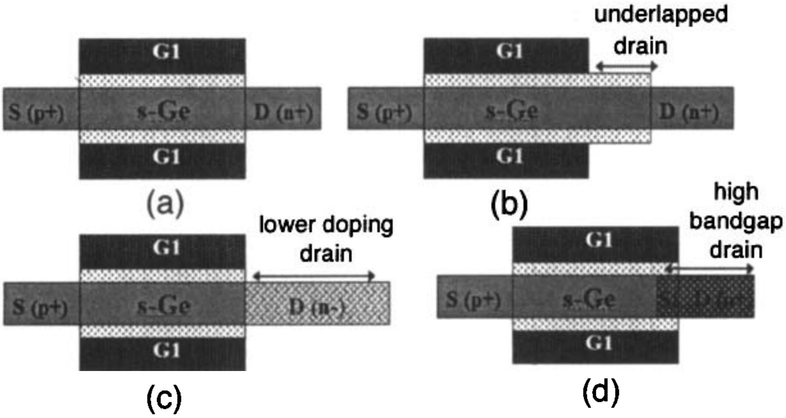


Figure 15. Symmetric s-Ge DG TFET structure (a) exhibits ambipolar characteristics. In order to achieve high I_{ON} and low I_{OFF}, asymmetry can be introduced into the TFET by (b) using an under-lapped drain; (c) using a lower doping on the drain side; or (d) using a lateral heterostructure of large bandgap material to reduce the tunneling currents from the drain in the off-state.[29]

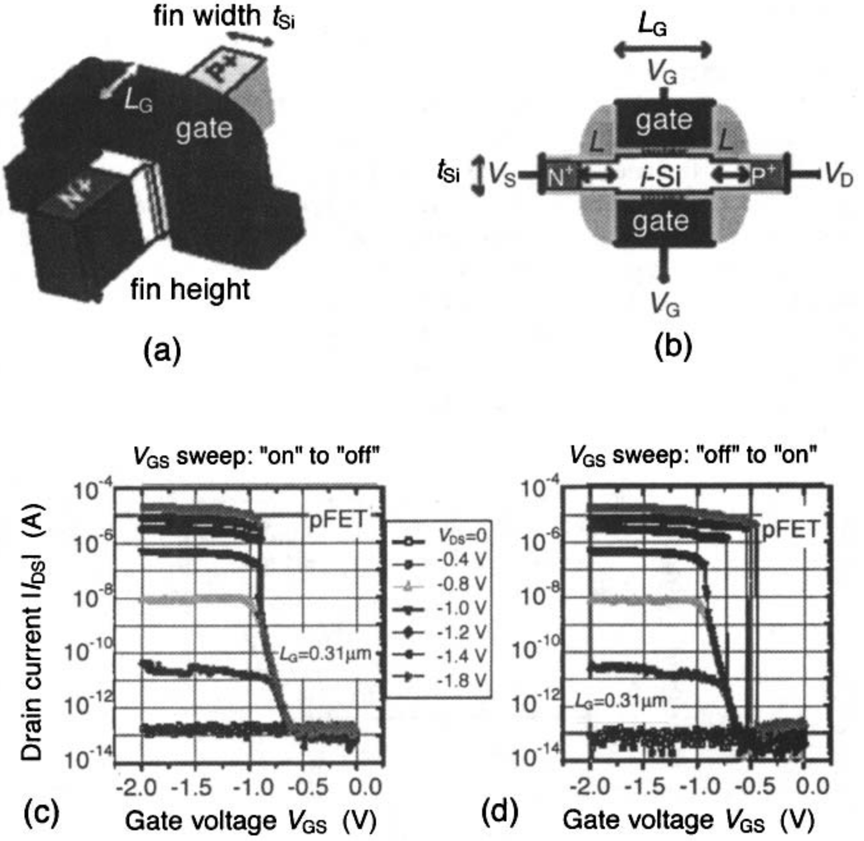


Figure 16. Schematic illustration (a) and cross-section (b) of the feedback FET; measured $I_D(V_G)$ characteristics of a short channel feedback FET operated as a p-channel device: (c) forward sweep, (d) reverse sweep, L_G = 0.31 μm, W = 0.2 μm.[30]

4. Advanced multi-gate SOI memories

It is becoming difficult to scale down memory devices. Indeed, traditional embedded DRAM requires a complicated stack capacitor or a deep trench capacitor in order to obtain a sufficient storage capacitance in smaller cells. This leads to more process steps and thus less process compatibility with logic devices. Capacitorless 1T-DRAMs or floating body cells have shown promising results. The operating principle is based on excess holes that can be generated either by impact ionization or by gate-induced drain leakage (GIDL) current in partially-depleted (PD) SOI MOSFETs. The GIDL current is due to band-to-band tunneling and occurs in accumulation, leading to a low drain current writing and reduced power consumption together with high-speed operation. However, conventional PD-SOI MOSFETs require high channel doping to suppress short-channel effects, which degrades the retention characteristics. In order to overcome this problem, DG fin-DRAMs, illustrated in Fig. 17, have been proposed showing superior memory characteristics.[31]

Conventional floating-gate flash memory also faces scaling problems due to the nonscaling gate-insulator stack and inefficient hot carrier injection processes at sub-50 nm gate dimensions. Back-floating gate flash memory overcomes these limitations by decoupling read and write operations and independent positioning and/or sizing of the storage element (back-floating gate) under the Si channel – see Fig. 18. The charge in the back gate affects the field and the potential at the bottom interface and thus changes the threshold voltage of the device. The back-floating gate is charged (programmed) by applying –10 V to the source, drain and front gate simultaneously, and the charges are removed from the back-floating gate (erased) with the same method but with a bias of +10 V.[32]

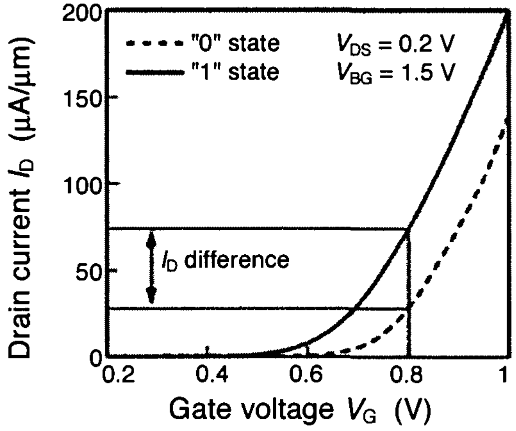


Figure 17. $I_D(V_G)$ characteristics of the capacitorless DG fin-DRAM with floating body.[31]

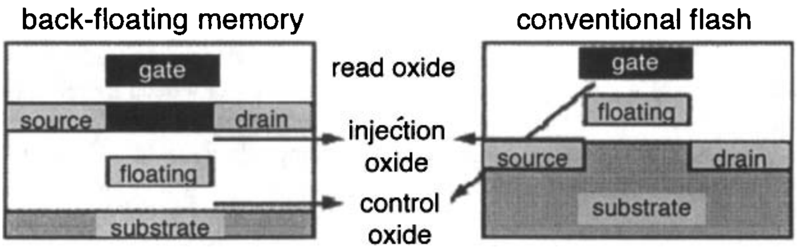


Figure 18. Cross-sectional diagrams of back-floating gate and conventional front-floating gate memories.

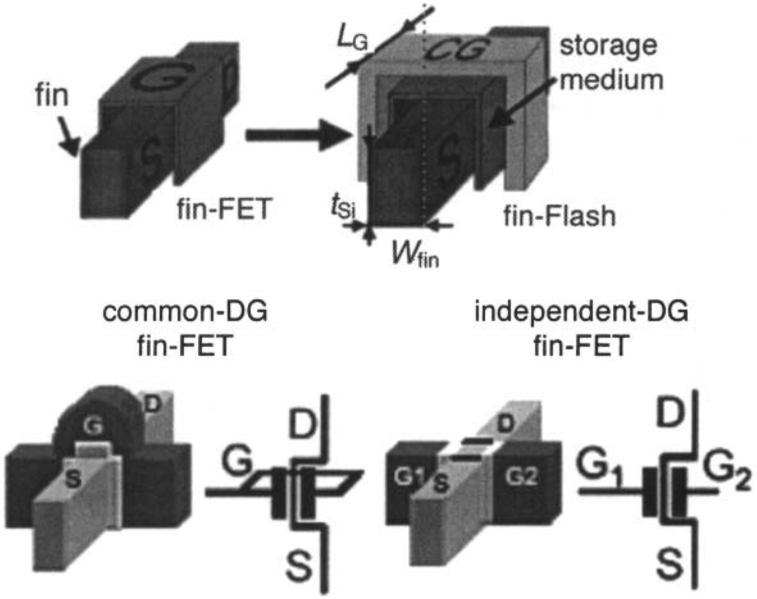


Figure 19. (a) Fin-FET used as a 1T-DRAM; (b) fin-Flash with poly-Si or SiN or nanocrystals; (c) schematic circuit diagrams of the common- and independent-DG fin-FETs.

Fin-Flash, SiN-based fin-Flash and nanocrystal fin-Flash on SOI, illustrated in Fig. 19, also offer many advantages: increased drive current and improved access time, reduced short-channel effects leading to integration capability down to at least 20 nm, higher numbers of stored electrons and thus less fluctuation problems, compatibility with future multi-gate CMOS, *etc.*[33] Furthermore, a reduction of leakage current and an enhancement of read and write noise margins have been successfully demonstrated in DG fin-FET SRAM cells by introducing the independent DG fin-FETs, shown in Fig. 19(c).[34]

Multi-gate SOI structures seem therefore to be very promising for extending the scalability and the performance of DRAM, SRAM and flash memories.

5. Conclusion

Silicon-on-insulator-based devices seem to be the best candidates for the ultimate integration of integrated circuits on silicon down to nm structures. An overview of the performance of nanoscale CMOS, emerging and beyond-CMOS nanodevices, based on innovative concepts, technologies and device architectures, has been presented in this chapter. The impact of alternative channel materials and multiple channels on the performance and physical mechanisms in multi-gate MOSFETs has been addressed. The interest of multi-gate emerging and beyond-CMOS nanodevices for long term applications, based on nanowires or small slope switch structure for ultra low power applications, has been presented. Finally, the flexibility of multi-gate SOI structures for boosting the scalability and performance of DRAM, SRAM and flash memories has been outlined. Multi-gate SOI platforms will make it possible to overcome a number of technological and physical challenges we are facing for ultimate CMOS and post-CMOS nanodevices in order to speed up technological innovation for the nanoelectronics of the next three decades.

Acknowledgments

This work was partially supported by the European Networks of Excellence NANOSIL (FP7) and SINANO (FP6) devoted to Silicon-based Nanodevices.

References

1. ITRS Roadmap: see *http://www.itrs.net/*
2. ENIAC strategic research agenda can be consulted at: *http://www.eniac.eu/web/SRA/local_index.php*
3. F. Balestra, "SOI devices", chapter in: *Wiley Encyclopedia of Electrical and Electronics Engineering*, New York: Wiley, 1999.
4. S. Cristoloveanu and F. Balestra, "Silicon on insulator: Technology and devices," chapter in H. Morkoc, ed., *Advanced Semiconductor and Organic Nano-Techniques*, New York: Academic Press, 2004, pp. 325–366.
5. F. Driussi, D. Esseni, L. Selmi, *et al.*, "Fabrication, characterization and modeling of strained SOI MOSFETs with very large effective mobility," *Proc. ESSDERC* (2007), pp. 315–318.
6. S. F. Feste, J. Knoch, S. Habicht, D. Buca, Q. T. Zhao, and S. Mantl, "Performance enhancement of uniaxially-tensile strained Si NW-nFETs fabricated by lateral strain relaxation of sSOI," *Proc. ULIS* (2009), pp. 109–112.
7. K. Uchida, R. Zednik, C.-H. Lu, *et al.*, "Experimental study of uniaxial and biaxial strain effects on carrier mobility in bulk and ultra-thin-body SOI MOSFETs," *Tech. Digest IEDM* (2004), pp. 229–232.
8. V. Barral, T. Poiroux, F. Andrieu, *et al.*, "Strained FDSOI CMOS technology scalability down to 2.5 nm film thickness and 18 nm gate length with a TiN/HfO_2 gate stack," *Tech. Digest IEDM* (2007), pp. 61–64.
9. G. Larrieu, E. Dubois, R. Valentin, *et al.*, "Low temperature implementation of dopant-segregated band-edge metallic S/D junctions in thin-body SOI p-MOSFETs," *Tech. Digest IEDM* (2007), pp. 147–150.
10. Z. Zhang, A. Qiu, R. Liu, M. Ostling, and S. L. Zhang, "Schottky-barrier height tuning by means of ion implantation into preformed silicide films followed by drive-in anneal," *IEEE Electron Dev. Lett.* **28**, 565 (2007).

11. O. Engström, B. Raeissi, S. Hall, and O. Buiu, "Navigation aids in the search for future high-κ dielectrics: Physical and electrical trends," *Solid State Electronics* **51**, 622 (2007).
12. F. Balestra, S. Cristoloveanu, M. Benachir, and J. Brini, "Double-gate silicon-on-insulator transistor with volume inversion: A new device with greatly enhanced performance," *IEEE Electron Dev. Lett.* **8**, 410 (1987).
13. S. Monfray, F. Boeuf, P. Coronel, G. Bidal, S. Denorme, and T. Skotnicki, "Silicon-on-nothing (SON) applications for low power technologies," *Proc. ICICDT* (2008), art. 4567232.
14. S. Eminente, D. Esseni, P. Palestri, C. Fiegna, L. Selmi, and E. Sangiorgi, "Enhanced ballisticity in nano-MOSFETs along the ITRS roadmap: A Monte Carlo study," *Tech. Digest IEDM* (2004), pp. 609–612.
15. M. Bescond, K. Néhari, J. L. Autran, N. Cavassilas, D. Munteanu, and M. Lannoo, "3D quantum modeling and simulation of multi-gate nanowire MOSFETs," *Tech. Digest IEDM* (2004), pp. 617–620.
16. J. Saint Martin, A. Bournel, and P. Dollfus, "Comparison of multiple-gate MOSFET architectures using Monte-Carlo simulation," *Solid State Electronics* **50**, 94 (2006).
17. A. Bansal, B. C. Paul, and K. Roy, "Impact of gate underlap on gate capacitance and gate tunneling current in 16 nm DGMOS devices," *Proc. IEEE Intern. SOI Conf.* (2004), pp. 94–95.
18. Q. Raphay, R. Clerc, G. Ghibaudo, and G. Pananakakis, "Impact of source-to-drain tunneling on the scalability of arbitrarily oriented alternative channel material nMOSFETs," *Solid State Electronics* **52**, 1474 (2008).
19. E.-J. Yoon, S. Y. Lee, and S. M. Kim, "Sub-30nm multi-bridge-channel MOSFET (MBCFET) with metal gate electrode for ultra high performance application," *Tech. Digest IEDM* (2004), pp. 627–630.
20. E. Bernard, T. Ernst, B. Guillaumot, *et al.*, "Novel integration process and performances analysis of low standby power (LSTP) 3D multi-channel MOSFET (MCFET) on SOI with metal/high-κ gate stack," *Tech. Digest VLSI Symp.* (2008), pp. 16–17.
21. O. Weber, O. Faynot, F. Andrieu, *et al.*, "High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding," *Tech. Digest IEDM* (2008), art. 4796663.
22. B. Cheng, S. Roy, A. R. Brown, C. Millar, and A. Asenov, "Evaluation of intrinsic parameter fluctuations on 45, 32 and 22nm technology node LP N-MOSFETs," *Proc. ESSDERC* (2008), pp. 47–50.
23. K. Moselund, P. Dobrosz, S. Olsen, *et al.*, "Bended gate-all-around nanowire MOSFET: A device with enhanced carrier mobility due to oxidation-induced tensile stress," *Tech. Digest IEDM* (2007), pp. 191–4.
24. C. Dupre, A. Hubert, S. Becu, *et al.*, "15 nm-diameter 3D stacked nanowires with independent gates operation: ΦFET," *Tech. Digest IEDM* (2008), art. 4796805.

25. T. Ernst, L. Duraffourg, C. Dupre, *et al.*, "Novel Si-based nanowire devices: Will they serve ultimate MOSFETs scaling or ultimate hybrid integration," *Tech. Digest IEDM* (2008), art. 4796804.
26. K. K. Bhuwalka, J. Schulze, and I. Eisele, "Performance enhancement of vertical tunnel field effect transistor with SiGe in the δp+ layer," *Japan. J. Appl. Phys.* **43**, 4073 (2004).
27. P.-F. Wang, K. Hilsenbeck, T. Nirschl, M. Oswald, C. Stepper, and M. Weis, "Complementary tunneling transistor for low power applications," *Solid State Electronics* **48**, 2281 (2004).
28. K. Boucart and A. M. Ionescu, "Double-gate tunnel FET with high-κ gate dielectric," *IEEE Trans. Elec. Dev.* **54**, 1725 (2007).
29. T. Krishnamohan, D. Kim, S. Raghunathan, and K. Saraswat, "Double-gate strained-Ge heterostructure tunneling FET (TFET) with record high drive currents and <60mV/dec subthreshold slope," *Tech. Digest IEDM* (2008), art. 4796839.
30. A. Padilla, C. W. Yeung, C. Shin, C. Hu, and T.-J. King Liu, "Feedback FET: A novel transistor exhibiting steep switching behavior at low bias voltages," *Tech. Digest IEDM* (2008), art. 4796643.
31. T. Tanaka, E. Yoshida, and T. Miyashita, "Scalability study on a capacitorless 1T-DRAM: from single-gate PD-SOI to double-gate FinDRAM," *Tech. Digest IEDM* (2004), pp. 191–92.
32. U. Avci, A. Kumar, and S. Tiwari, "Back-floating gate nonvolatile memory," *Proc. IEEE Intern. SOI Conf.* (2004), pp. 133–35.
33. E. Nowak, M. Bocquet, L. Perniola, *et al.*, "New physical model for ultra-scaled 3D nitride-trapping non volatile memories," *Tech. Digest IEDM* (2008), art. 4796750.
34. K. Endo, S.-I. O'uchi, Y. Ishikawa, *et al.*, "Enhancing SRAM cell performance by using independent double-gate FinFET," *Tech. Digest IEDM* (2008), art. 4796833.

Device Proposals Beyond Silicon CMOS

P. M. Solomon
IBM, SRDC, T.J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A.

1. Introduction

As the end of the silicon evolutionary path nears, alternative devices are being proposed on an urgent basis. Such devices involve different materials such as carbon, III-V semiconductors, *etc*., different geometries, such as nanotubes, nanowires and graphenoid sheets, and different operating principles involving collective phenomena such as coherent tunneling and ferroelectricity and density of states engineering for band-to-band tunneling FETs. All add to a weird device menagerie that needs some sorting out. These device proposals are mostly not new, but they are enabled by the march of technology and the apparent need for a device that breaks the inflexible switching-energy *vs*. performance limit of silicon CMOS technology. Meanwhile the goalposts shift continually with the evolution of CMOS technology and system design. Here I will attempt to describe and evaluate the most promising to the most outlandish of these devices in terms of future needs for large scale computation.

2. Industry view

End of CMOS scenarios and successor technologies to CMOS have been the fascination of the semiconductor industry for at least the past decade. Numerous project initiatives and focus centers have provided support into research on alternative devices, circuits and architectures, most notably as part of the Semiconductor Research Corporation's (www.src.org) Nanoelectronics Research Initiative (NRI) and Focus Center Research Program (FCRP). So far the lack of success has been notable except perhaps for the RSFQ logic family,[1] an ultra-high speed superconducting technology operating at cryogenic temperatures. This logic family has been dropped from the ITRS (2007) menu because working prototypes had been demonstrated but the market had not materialized. The industry view is encapsulated in the chart[2] reproduced in Fig. 1 with the "state variable" shown at the bottom and proceeding to increasingly higher levels of implementation toward the top (although a strictly 1:1 progression is not implied). A primary motivation exemplified by this chart is to replace charge with some other variable (polarization, spin, phase, *etc*.), and the rationale is that a non-charge based state variable may lead to a smaller switching energy since it avoids the electrostatic

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

energy associated with charging the gate capacitance of an FET. Thus we see spintronics features quite prominently in the research efforts as well as propagation of electric or magnetic polarization in quantum cellular automata (QCA) type of effects.

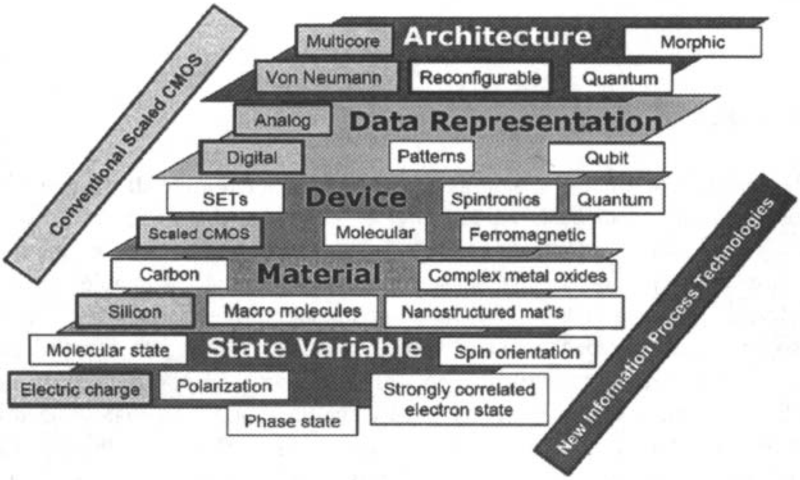


Figure 1. Chart from 2007 International Technology Roadmap, Emerging Research Devices[2] showing the emphasis on other state variables besides charge for the new information process technologies. [Reproduced by permission from the Semiconductory Industry Association, International SEMATECH: Austin, TX (2007)].

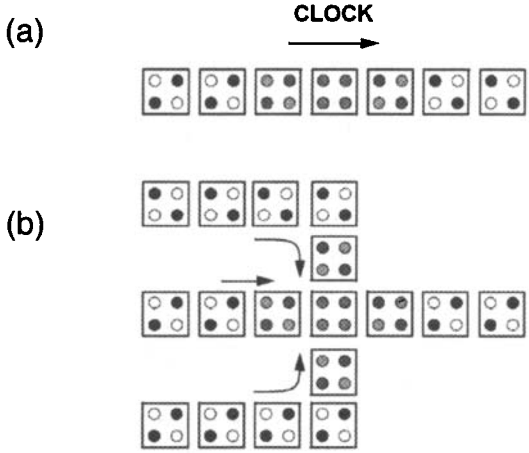


Figure 2. Quantum cellular automaton (QCA) principle, after Lent:[3] (a) clock-driven polarization wave propagation along a chain of QCA gates; (b) majority logic gate where the bottom two inputs determine the output state.

Setting aside quantum computation, which is really in a class by itself and will not be discussed here, the proposed solutions fit rather poorly into Von Neumann type architectures, thus research into compatible architectures is integral to assessing the place of some of the more exotic proposals. Much excellent work has been done in exploring these various avenues, and many significant advances have been made, some of which will be discussed below. Still, so far it seems that the further one strays from the CMOS path (left edge of the chart) the less viable the proposals seem to be. Why is it that CMOS seems to have such an unbreakable monopoly?

While competitors so far have been unable to assault the well-nigh impregnable CMOS fortress, CMOS and silicon technology is changing in ways that perhaps can give potential competitors a foothold. With silicon nearing the end of its scaling potential, many new solutions are being tried involving an expanding materials inventory. Silicon itself, apart from its role of supporting substrate, is only one layer among many for silicon-on-insulator realizations. Novel self-assembly and hybrid three-dimensional (3D) integration schemes allow the incorporation of other technologies into the silicon mix. Thus the path towards incorporating a novel logic or memory technology is becoming easier.

3. Other state variables

Computational state variables are simply the physical attributes of a system that carry computational information. The motivation to replace the presently used state variable (voltage, misleadingly called charge) is that the bulk of the energy to switch states (voltages) is used to charge circuit nodes comprising of internal device capacitances, parasitic inter-electrode capacitances and interconnects. Note that the logic is propagated electromagnetically from circuit to circuit at speeds approaching the speed of light. It is the nonlocal character of the voltage distribution that costs so much in terms of electrostatic energy. Nanoscale CMOS at the 22 nm node of the ITRS Roadmap has internal switching energies of $\sim 10^{-18}$ J for a minimum sized device but $\sim 10^{-16}$ J when including interconnects.

There is some confusion, as evidenced in Fig. 1, between the above definition and the use of "state variable", such as charge, to describe the internal state of a logic-switching device. In conventional devices this relationship is not one-to-one, since the internal charge is a function of device size, circuit design, *etc*. An abrupt change of state is desired, controlled by a small change of the terminal voltages, which is why collective and strongly correlated states are being sought in new devices.

When an internal state variable is used to directly represent digital bits, it is called a token. Information processing occurs by physically passing the token from device to device. This system is attractive because of the isomorphism between the state variable and the logic state heralds perhaps a greater energy efficiency compared to the conventional approach and perhaps a more robust representation compared to the arbitrary nature of voltage representation.

The token-carrying logic is exemplified by the QCA[3] approach, as shown in Fig. 2 where the states are represented by two diagonal alignments of internal polarization. The states are separated in energy by several *kT*, yet transitions can be propagated from cell-to-cell by means of a traveling wave generated by external clock electrodes that adiabatically facilitate the dipole rotation by compensating for the internal potential so that the cell polarization may be driven by weak neighboring fields. An instructive example was recently given for a magnetic QCA.[4,5] Calculations of energy dissipation for adiabatic transitions (slowly variable field) for flipping of electric and magnetic polarizations[3,4] and spins[6] show that it may be very small, much less than *kT*. Thus far, efforts in the field of token-passing logic have focused naturally on demonstrations of proof of concept, some with notable success[3] but a set of more fundamental questions remain.

Compared to electromagnetic propagation, the speed of propagation of the tokens is much smaller, typically by a factor of $\sim 10^3$. Also, logic interactions are predominantly via neighboring cells (cellular automaton). This gives rise to the following set of questions which to date have not been satisfactorily addressed: (i) cellular automata have inherent limitations and inefficiencies in implementing general purpose logic, which would result in some penalty factor vis-à-vis CMOS; (ii) many state transitions are utilized for communication rather than logic, which would result in further penalties; (iii) how does the communication penalty associated with token passing limit applications? (iv) the reactive power for such a system will be very large (creating the propagator fields) so that very high *Q* clock power supplies will be needed to maintain efficiency, but at present there are no solutions for this.

A family of proposed devices uses state variables other than charge to modulate their switching characteristics while still using external voltage and current for logic propagation. These include the Das–Datta spin transistor[7,8] (see Fig. 3) where the gate modulates the spin lifetime, the Mott transition,[9,10] quantum correlated states,[11] quantum interference devices, *etc.* The device itself needs to be a two-way transducer, converting from the terminal voltages to the internal state variable and back again. The advantages of changing state variable, for instance a claimed lower operating voltage for the spin-transistor, have to offset losses in the transducer chain, and this has been difficult to achieve.

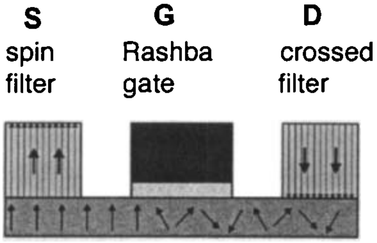


Figure 3. Spin transistor, after Datta and Das,[7] with crossed polarizers on source and drain. Enhanced spin-orbit coupling due to the gate field promotes rapid injected spin thermalization and thence transmission of a portion into the drain.

4. Scaling

Capacitance scales inversely with distance and density with area, so that scaling capability has always been an important and desirable characteristic for all new device proposals. The electrostatic QCA, for instance, has the ability to scale all the way down to molecular dimensions. Scaling has, of course, been extensively discussed elsewhere, but here we will briefly touch on aspects of new devices which make them interesting from the scaling perspective. Even the more conventional technologies are exploring quantum-confined geometries such as nanowires and nanotubes. These confer scaling advantages as a result of reduced dimensionality much like those conferred on 1D quantum confined lasers. For instance, as illustrated in Fig. 4, 1D quantum confinement collapses the transverse density of states into a single quantum number (not counting degeneracy), so that the current-carrying capacity of the quantum channel is independent of device cross-section. Thus carbon nanotubes can be scaled below 1 nm diameter and are still capable of carrying twice (for band degeneracy) the full quantum of conductance $2e^2/h$ times supply voltage worth of current, typically ~20 μA. Some shibboleths may fall by the wayside, such as the need for the gate to always be in close proximity to the channel of an FET. As illustrated in Fig. 5 it may be possible use highly polarizable materials to transmit potentials into small devices[12] or even molecules,[13] to control their switching. Another exciting possibility is to use collective effects to suppress single-electron tunneling and reduce lateral dimensions. Some work on oxide semiconductors, while still open to interpretation, has shown control of device properties on an extremely small scale.[14]

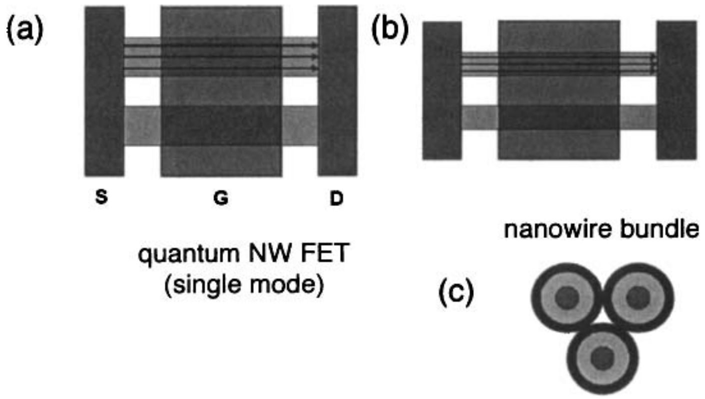


Figure 4. Nanowire scaling: shrinking width of nanowire (a) to (b) compresses the same amount of active charge into a smaller width therefore increasing the ratio of active to parasitic capacitance; (c) reduction of parasitic capacitance by nanowire bundling.

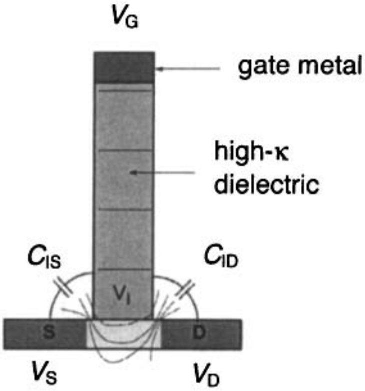


Figure 5. Field effect transistor with a very high-permittivity gate dielectric where the gate dielectric thickness may be much larger than the channel length, C_{IS} and C_{ID} are the gate fringing capacitances to source and drain respectively.

5. Beating *kT*(/*e*)

At this point in the evolution of integrated computing the overwhelming concern is power reduction or, in terms of individual device properties, the energy stored or dissipated per switching event. With today's numbers of transistors per chip at ~1 billion, total power constrained to below ~1 W and frequencies in excess of 1 GHz, average switching energies need to be below 10^{-17} J per transistor, and these demands will increase exponentially with time. As Table 1 shows, there is a distinction between switching energy kT and the energy per electron kT/e. The former places an absolute limit on the energy cost of nonreversible computation,[15] while the latter places a restriction on power supply voltage as applied to electron-barrier controlled devices such as FETs. That is, for a given I_{ON}/I_{OFF} current ratio r, a switching voltage $V = (kT/e)\ln r$ is required. This is the famous "60 mV/decade" subthreshold slope problem, and much research effort, funded by a dedicated government program, is devoted to finding devices with steeper that 60mV/decade slopes. This voltage requirement, coupled with the fact that the capacitance per unit length of the interconnects is a constant ~1 times the permittivity of the dielectric, means that switching energy, $\frac{1}{2}CV^2$, is rather insensitive to technology changes[16] depending only on the general length scale. To reduce switching energies therefore requires device, circuit and architectural innovation in addition to scaling. Here we will focus on device innovation. As discussed in a previous paper in this series[17] carbon-nanotube FETs may enable improved switching energies compared to CMOS simply by virtue of their higher performance, therefore offering a better power-performance trade-off, but to go beyond this requires devices operating on different principles.

Table 1 lists various "workaround" solutions, both to the 60 mV/decade problem and also to the fundamental "kT limit". The latter limit only applies to irreversible operations, *i.e.* all conventional computation, but the token-passing kind of logic including QCA with electric or magnetic dipoles, and spintronics generally postulate a reversible mode of computation where the clock is varied adiabatically and much of the energy can be recovered (notwithstanding the questions we posed earlier).

It should be noted that CMOS logic could, in principle, also be run in a reversible or partly reversible manner,[18] but the technical obstacles to achieving significant energy savings are daunting.

The kT and kT/e limits may be circumvented simply by reducing temperature, and this approach has a long history,[19] but in the end refrigerator inefficiencies and the Carnot factor have to be taken into account. Much has been made of the fact that electronic spin interacts weakly with the thermal bath and therefore may achieve lower energy dissipation.[6] An illustrative scheme is shown in Fig. 6. Spin-polarized electrons are injected into a nonmagnetic semiconductor via a polarizer. In the semiconductor the spins may be flipped by weak fields involving voltages of $\ll kT/e$ and may interact with each other to do logic. In reality what we have here is a refrigerator, since the injected spins have a super-cooled distribution in the zero-magnetic gap semiconductor.[20] Logic can be performed as long as execution times are much shorter than the thermalization time, and the electrons may be extracted (read) via a similar polarizer at no energy cost. However any electrons where the spin has been flipped, either intentionally or through interaction with the thermal bath, will have to be extracted via a complementary polarizer, but to prevent back-injection of the opposite spin polarization and hence contamination of the distribution, this polarizer has to be biased at a voltage of several kT/e with respect to the original injector. From this it is clear that the Carnot penalty is paid back (at least) during this extraction process. It may be justifiably argued that this constitutes a very compact and efficient refrigerator, but remember that this scheme is incomplete since only one degree of freedom (spin) has been cooled this way.

Energy	**Voltage**	**Distance**
kT	kT/e	h/p
Thermodynamic constraint for irreversible computing	Consequence of charge on single electron	Consequence of mass of single electron
Workaround		
Reversible computing	▪ energy filtering ▪ spin filtering ▪ collective effects	▪ m^* engineering ▪ collective effects

Table 1. Physical constraints.

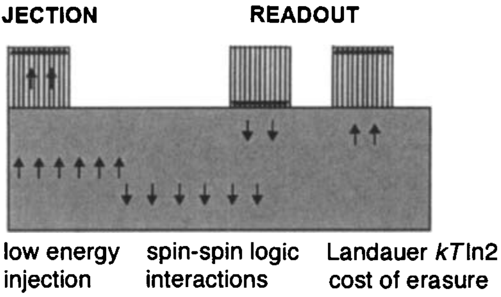

INJECTION
READOUT
low energy injection
spin-spin logic interactions
Landauer *kT*ln2 cost of erasure

Figure 6. Device with spin polarizers for injection and readout. The spin transport medium has zero bandgap for spin-spin interactions. The readout electrodes have to be biased at a sufficient potential to prevent back injection.

Practically the *kT/e* limit is more important than the *kT* limit since voltage and interconnect capacitance place switching energies today above 10^4 *kT* and this places emphasis on strategies to increase the sharpness of the switching transition. Two such strategies, listed in Table 1, are energy filtering and collective effects. These will be dealt with in the following sections.

6. Energy filtering

The principle of energy filtering is outlined in a previous FTM book chapter[21] and indeed can be traced back to the Esaki diode. The principle, as applied to an *n*-FET, is illustrated in Fig. 7. When the FET is in the "on" state, electrons originating in the valence band of the *p*-type source tunnel into the conduction-band of the channel. The device is turned off by raising the conduction-band edge in the channel above the valence band edge in the source. One can say that the Fermi tail of the electrons in the valence band has been cut off above the valence band edge permitting a steeper than 60mV/decade subthreshold slope. While the principle is clear, achieving a steep slope in practice has been difficult and so far has only been demonstrated unambiguously in a carbon nanotube geometry,[22] and only at low currents, unsuitable for high-speed devices. Band-to-band tunneling transmission coefficients decrease exponentially with $m_r^{1/2}$ and $E_G^{3/2}$ where m_r is the reduced effective mass for tunneling and E_G is the band gap. Thus materials with small m_r and E_G are desired. A direct bandgap is also important to achieve large current levels. Materials such as carbon nanotubes, graphene ribbons,[23] III-V semiconductors and Ge (with an almost direct bandgap) have the required properties.

Incorporation of heterojunctions,[24] as shown in Fig. 7(b), can provide the correct band line-up for the desired tunneling while suppressing tunneling where it is not wanted, such as from the drain into the channel. Nanowires (or nanotubes), with wrap-around gates, are the preferred geometry since this provides an intimate electrostatic control of the tunnel junction by the gate.

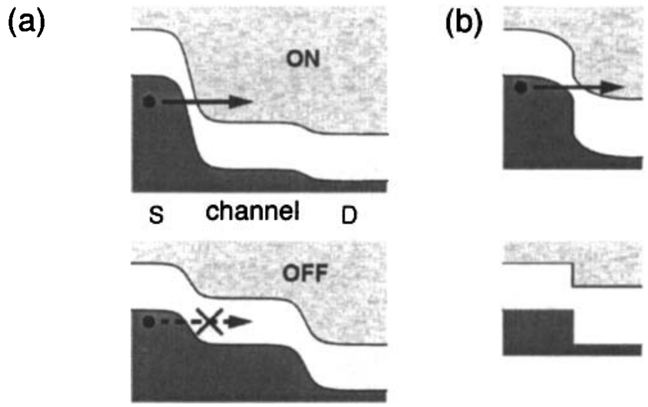


Figure 7. Operating principle of a tunnel field-effect transistor, after Appenzeller *et al.*:[22] (a) the bands are crossed in the "on" state and uncrossed by the gate voltage in the "off" state; (b) a staggered gap heterojunction achieves a smaller bandgap for tunneling while allowing a larger bandgap in the source and channel regions.

7. Band structure engineering

While conventional device design has assumed the band structure as a given (*e.g.* bulk silicon), quantum effects on the nanoscale can alter band energies. This may be exploited, in the case of graphene, to make nanoribbons with controlled bandgaps as in proposals for the tunnel FET.[23] For graphene nanoribbons of a certain type the dispersion of edge states (see Fig. 8) may be controlled by a lateral field and this dispersion-modulation has been proposed as a new way of modulating the current,[25] since transport can occur only when there is finite dispersion. Similarly, other proposals exploit the ability of a perpendicular field to modulate the bandgap of bilayer graphene.[26]

8. Collective effects

Going back to Table 1, the kT/e potential can simply be replaced by kT/ne where n is much larger than unity. In other words, collections of correlated particles still have mean thermal energies of $\sim kT$ but a much lower electric potential than kT/e (as seen in Table 1, increased n also offers potential scaling benefits since it reduces h/p). Collective effects are being evoked to extend devices beyond ultimate CMOS, where dimensions are on the 10 nm scale. A legitimate question therefore is how do the collective effects withstand scaling where n decreases perhaps as fast as the cube of the dimension. In the case of ferromagnetism, experiments and theory indicate that scaling down to ~5nm is possible,[27] and experiments on semiconducting oxides[14] show effects persisting on the ~2 nm distance scale.

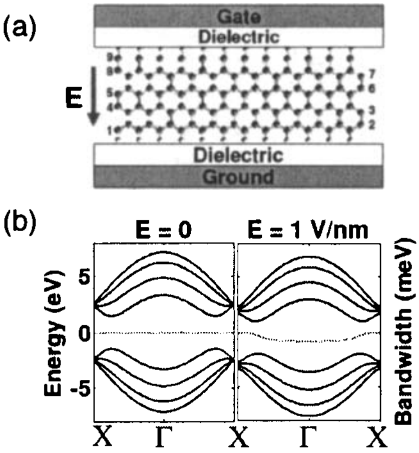


Figure 8. Bandwidth modulation FET, after Raza:[25] (a) lateral gate applied to a graphene nanoribbon; (b) gate field breaks the symmetry and adduces a finite bandwidth to the midgap edge states.

This can be exploited in switching ferromagnetic of ferroelectric domains, in the Mott metal-insulator transition, in correlated electron condensations such as bi-layer graphene[28] or semiconducting oxides[29] and in correlated tunneling. In addition to the number effect, there are also quantum exchange interactions that reduce potential energies and can give rise to a "negative capacitance",[30] and by inference to switching behavior. As is seen from the above list, the field is rich and just beginning to be explored. Here we will cite just two examples of current research interest.

Ferroelectricity leads to hysteretic charge *vs.* voltage characteristics (analogous to the well known hysteretic loops of ferromagnetism). This is being exploited commercially for memory applications and it has also been proposed for logic devices as a way of increasing the subthreshold slope.[31] It is proposed that by combining the negative capacitance of an unstable ferroelectric state with the positive gate capacitance an FET, that the system may be made marginally stable with the internal gain resulting in an almost vertical off-on transition – see Fig. 9. The viability of this idea involves many questions concerning domain formation, gain per unit volume, speed, *etc.* Similar proposals are in place for a purely electronic negative capacitance resulting from strong quantum mechanical exchange effects,[30] as have been seen in oxide semiconductors. Ferromagnetism itself may be used for logic as in the case of magnetic QCA. Also, ferromagnetic spin-wave logic[32] is being explored. Research is also being done into multiferroics, especially composite coupled systems,[33] where electric fields may control ferroelectric properties and *vice versa.*

The other example involves inducing correlated tunneling across two graphene layers[28] separated by an insulator, as shown in Fig. 10. When the "nesting" condition is achieved and the electron and hole Fermi surfaces on the two sides are matched, correlated tunneling can occur and the resistance between the two layers is reduced from an insulator-like value to just the quantum of conductance. The remarkable prediction is that the conductance in the "on" state is just the quantum of conductance $2e^2/h$ and does not depend on the insulator thickness, even though uncorrelated tunneling current is thus reduced exponentially. This state has not been experimentally confirmed, but a low-voltage device has been proposed by analogy to the superconducting Josephson junction.[34]

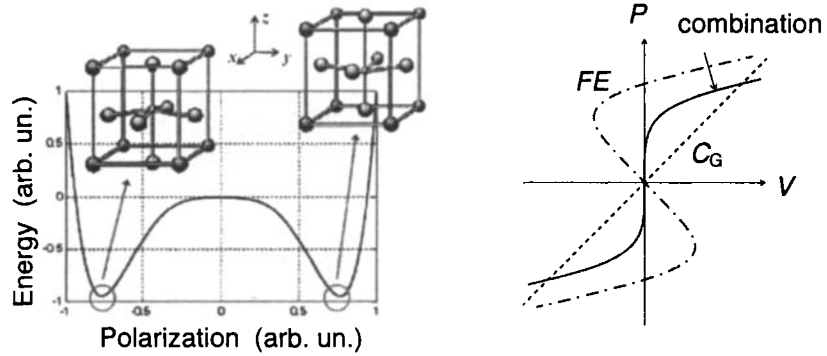


Figure 9. Ferroelectric element (a) in series with the gate of an FET; (b) the ferroelectric response (derivative of (a), highly idealized) is added to the response of the series gate capacitor C_G designed to maximize the change in polarization following small changes in gate voltage. After Ref. 31.

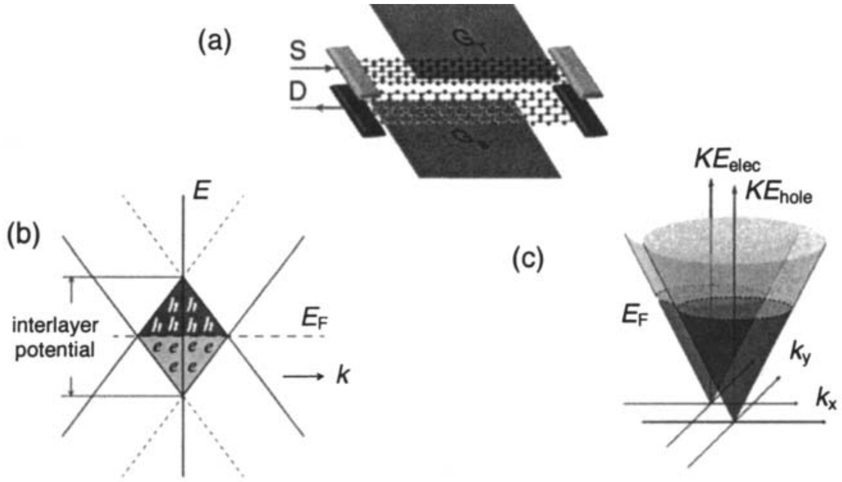


Figure 10. Bilayer pseudo-spin FET, after Banerjee:[34] (a) two graphene sheets separated by an insulator and gated on top and bottom; (b) bands of the two layers shifted by the gate-induced potential into the "nesting" condition where the electron and hole Fermi surfaces coincide (c), for correlated tunneling between the layers.

9. Outlook

In the above treatment we have tried to convey a flavor of the many approaches used and avenues being investigated to come up with a future device that is better, mainly in terms of power dissipation, than CMOS. Much has been left out and much worthy work left unmentioned, due to lack of space and lack of personal familiarity. For this I apologize. Most of the approaches have not yet resulted in working demonstrations, let alone CMOS-competitive ones. This does not in anyway diminish the quality and importance of this work, since truly a new frontier is being explored and only those in the future, looking back, will be able to evaluate the fruits of today's efforts.

Acknowledgments

I wish to acknowledge the help of the following in the form of discussions and material supplied: Siyuranga Koswatta, Jeff Welser, and Steven Koester.

References

1. P. Bunyk, K. Likharev, and D. Zinoviev, "RSFQ technology: Physics and devices," *Intern. J. High Speed Electron. Syst.* **11**, 257 (2001).
2. 2007 International Technology Roadmap, Emerging Research Devices, *www.itrs.net/Links/2007ITRS/2007_Chapters/2007_ERD.pdf*
3. C. S. Lent and B. Isaksen, "Clocked molecular quantum-dot cellular automata," *IEEE Trans. Electron Dev.* **50**, 1890 (2003).
4. G. Csba, P. Lugli and W. Porod, "Power dissipation in nanomagnetic logic devices," *Digest 4th IEEE Conf. Nanotechnol.* (2004), p. 346.
5. B. Behin-Aein, S. Salahuddin, and S. Datta, "Switching energy of ferromagnetic logic bits," *IEEE Trans Nanotechnol.* **8**, 505 (2009).
6. D. E. Nikonov, G. I. Bourianoff, and P. A. Gargini, "Power dissipation in spintronics devices away from thermal equilibrium," *J. Supercond. Novel Magn.* **19**, 497 (2006).
7. S. Datta and B. Das, "Electronic analog of the electro-optic modulator," *Appl. Phys. Lett.* **56**, 665 (1990).
8. K. C. Hall and M. E. Flatté, "Performance of a spin-based insulated gate field effect transistor," *Appl. Phys. Lett.* **88**, 162503 (2006).
9. C. Zhou, D. M. Newns, J. A. Misewich, and P. C. Pattnaik, "A field-effect transistor based on the Mott transition in a molecular layer," *Appl. Phys. Lett.* **70**, 598 (1997).
10. C. Ko and S. Ramanathan, "Observation of electric field-assisted phase transition in thin film vanadium oxide in a metal-oxide-semiconductor device geometry," *Appl. Phys. Lett.* **93**, 252101 (2008).

11. S. Thiel, G. Hammerl, A. Schmehl, C. W. Schneider, J. Mannhart, "Tunable quasi–two-dimensional electron gases in oxide heterostructures," *Science* **313**, 1942 (2006).
12. W. Haensch, E. J. Nowak, R. H. Dennard, *et al.*, "Silicon CMOS devices beyond scaling," *IBM J. Res. Develop.* **50**, 339 (2006).
13. N. D. Lang and P. M. Solomon, "Charge control in a model biphenyl molecular transistor," *Nano Lett.* **5**, 921 (2005).
14. C. Cen, S. Thiel, G. Hammerl, *et al.*, "Nanoscale control of an interfacial metal–insulator transition at room temperature," *Nature Mater.* **7**, 295 (2008).
15. R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Develop.* **3**, 183 (1961).
16. P. M. Solomon, "Comparison of semiconductor devices for high-speed logic," *Proc. IEEE* **70**, 489 (1982).
17. P. M. Solomon, "Carbon-nanotube solutions for the post-CMOS-scaling world," chapter in: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: Up the Nano Creek*, New York: Wiley, 2007.
18. M. P. Frank, "Reversible computing and truly adiabatic circuits: The next great challenge for digital engineering," 5th IEEE Dallas Circ. Syst. Workshop (DCAS-06), Richardson, TX, 2006.
19. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. Le Blanc, "Design of ion-implanted MOSFETs with small physical dimensions," *IEEE J. Sol. State Circ.* **9**, 256 (1974).
20. We are ignoring, for pedagogical purposes, the non-spin components of the electron's energy, which, of course, still retain their original thermal distributions.
21. P. M. Solomon and C. R. Kagan, "Understanding molecular transistors," chapter in: S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends In Microelectronics: The Nano Millennium*, New York: Wiley, 2004.
22. J. Appenzeller, Y.-M. Lin, J. Knoch, and Ph. Avouris, "Band-to-band tunneling in carbon nanotube field-effect transistors," *Phys. Rev. Lett.* **93**, 196805, (2004).
23. Q. Zhang, T. Fang, H. Xing, A. Seabaugh, and D. Jena, "Graphene nanoribbon tunnel transistors," *IEEE Electron. Dev. Lett.* **29**, 1344 (2008).
24. J. Knoch, M. T. Bjoerk, H. Riel, H. Schmid, M. Enderlein, H. Ghoneim and W. Riess, "Transport in heterostructure nanowire tunneling FETs," *2nd Intern. Conf. One-Dimensional Nanomater.*, Malmö, 2007.
25. H. Raza, T. Z. Raza, T.-H. Hou, and E. C. Kan, "On the possibility of an electronic-structure modulation transistor," *arXiv:0812.0123v2* and *v1* (2009).
26. G. Fiori and G. Iannaccone, "On the possibility of tunable-gap bilayer graphene FET," *IEEE Electron. Dev. Lett.* **30**, 261 (2009).
27. H. J. Richter, A. Y. Dobin, R. T. Lynch, *et al.*, "Recording potential of bit-patterned media," *Appl. Phys. Lett.* **88**, 222512 (2006).
28. J.-J. Su and A. H. MacDonald, "How to make a bilayer exciton condensate flow," *Nature Phys.* **4**, 799 (2008).

29. S. Thiel, G. Hammerl, A. Schmehl, C. W. Schneider, and J. Mannhart, "Tunable quasi-two-dimensional electron gases in oxide heterostructures," *Science* **313**, 1942 (2006).
30. T. Kopp and J. Mannhart, "Calculation of the capacitances of conductors – Perspectives for the optimization of electronic devices," *J. Appl. Phys.* **106**, 1 (2009).
31. S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Lett.* **8**, 405 (2008).
32. M. P. Kostylev, A. A. Serga, T. Schneider, B. Leven, and B. Hillebrands, "Spin-wave logical gates," *Appl. Phys. Lett.* **87**, 153501 (2005).
33. Y.-H. Chu, L. W. Martin, M. B. Holcomb, *et al.*, "Electric-field control of local ferromagnetism using a magnetoelectric multiferroic," *Nature Mater.* **7**, 478 (2008).
34. S. K. Banerjee, L. F. Register, E. Tutuc, D. Reddy, and A. H. MacDonald, "Bilayer pseudo-spin field-effect transistor (BiSFET): A proposed new logic device," *IEEE Electron. Dev. Lett.* **30**, 158 (2009).

GeOI as a Platform for Ultimate Devices

W. Van Den Daele, S. Cristoloveanu
IMEP-LAHC/Minatec, Grenoble-INP, 3 parvis Louis Néel, Grenoble, France

E. Augendre, C. Le Royer, J.-F. Damlencourt
CEA–LETI/Minatec, F38014, Grenoble, France

D. Kazazis and A. Zaslavsky
Div. of Engineering, Brown University, Providence, RI 02912, U.S.A.

1. Introduction

The International Roadmap for Semiconductors predicts increasing device and circuit performance through the reduction of the physical MOSFET gate length for years to come.[1] At the same time, the supply voltage V_{DD} will also continue to decrease, whereas the dynamic power consumption has to be limited to around 100 W/cm^2 for the future Si technology nodes (< 45 nm). Hence, a strong enhancement of the transport properties is the key point for reliable future nodes. The on-state current I_{ON} of the MOSFETs is intrinsically linked and limited by the active semiconductor material, in particular for short-channel technologies where ballistic transport appears. Besides the incremental technological booster solutions for silicon CMOS (source-drain engineering, SOI, strained SOI, *etc.*), there is a strong research effort to conceive of new architectures, like the 3D stacking,[2] or to replace silicon-based channels with higher mobility materials.

Germanium had been used decades ago to create the first generation of field effect and bipolar transistors. Unfortunately, the poor quality of its native oxide (GeO_2) made it unsuitable for VLSI. However, the recent development of high-κ dielectrics and metal gates has rekindled interest in Ge as a promising candidate to replace Si for advanced CMOS technologies.[3] The main advantages of Ge are: enhanced transport properties, especially for holes; a better match between electron and hole mobilities; and an excellent lattice match with GaAs (suitable for Ge and GaAs co-integration). The lower bandgap of Ge, while deleterious for leakage currents in standard MOSFETs, may also prove advantageous for advanced devices. Similarly to SOI, the germanium-on-insulator (GeOI) platform is beneficial for Ge MOSFET development through the improvement of the electrostatic control (essential for nanometric technologies) and the reduction of the Ge material consumption.[4, 5]

Recently, high-performance 65 nm Ge PMOS devices with pockets, halos and NiGe metal source/drain metallization have been demonstrated: very good I_{ON} = 478 μA/μm and low I_{OFF} = 0.9 μA/μm were achieved.[6] The corresponding NMOS

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

devices, on the other hand, have been less impressive. It is speculated that both the poor NMOS performance and the undesirable positive threshold voltage V_T shift in Ge PMOS arise from the presence of unpassivated dangling-bond (DB) states located close to valence band edge,[7,8] or from negatively charged interstitial hydrogen H^-.[9] Recently, alternative gate dielectrics technologies, such as GeO_2 pre-gate passivation with HfO_2 followed by fluorine post-gate treatment,[10] and GeO_2/Al_2O_3 stacking,[11] have shown promise of improving on standard SiO_x/HfO_2 approaches. If the Ge gate insulator passivation issue is resolved, GeOI will have a realistic prospect of mainstream technological insertion.

In this chapter, we review the recent progress on GeOI substrate manufacturing using both principal fabrication techniques, Smart-Cut^{TM} and Ge enrichment, as well as variants like strained GeOI, Si-Ge co-integration, *etc.* We then discuss GeOI material characterization by the pseudo-MOS (Ψ-MOS) technique. The chapter concludes with a discussion of GeOI as platform for high-performance PMOS, as well as alternative devices, such as tunneling FETs (TFETs) and nanowires.

2. GeOI substrate engineering: Smart-Cut^{TM} and Ge enrichment

The Smart-Cut™ technology can be applied to fabricate GeOI substrates. The Ge donor wafers can either be bulk substrates,[12] or Si substrates capped with epitaxial Ge.[13] The process sequence is illustrated in Fig. 1(a). After cleaning and chemical vapor oxide deposition (at 400 °C), Ge donor substrates are implanted with hydrogen (a few $10^{15}/cm^2$ at 50–100 keV) and prepared for hydrophilic bonding onto a silicon base wafer at room temperature. After subsequent thermal processing, the bonded pair splits, producing a GeOI substrate and a Ge donor wafer that can be reused. Finally, polishing and annealing steps complete the fabrication of the GeOI substrate, delivering a smooth Ge film – see Fig. 1(b).

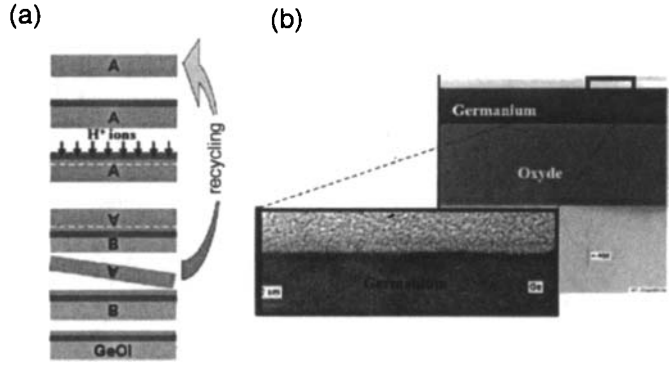


Figure 1. (a) Process sequence for the fabrication of GeOI via the Smart-Cut™ technology. Donor substrate A can be bulk Ge or epitaxially grown Ge. Substrate B is a Si handle wafer; (b) HR-TEM cross-section of the final GeOI substrate.

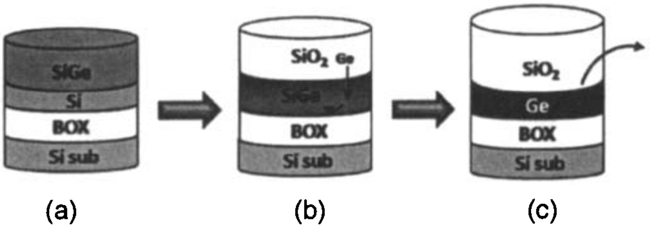


Figure 2. GeOI fabrication process by Ge condensation via oxidation: (a) SiGe layer is grown epitaxially on an SOI wafer; (b) oxidation of SGOI; (c) Ge is completely condensed and the top oxide is removed.

Depending on the Ge donor substrate nature and on the Ge/oxide interface, several GeOI substrate variants can be produced, like GeO_xN_y interface or compressively strained Ge layer on insulator.

The Ge enrichment (also called condensation) technique consists of germanium enrichment of a SiGe layer during oxidation. The first step of the condensation technique, illustrated in Fig. 2, is based on the growth of a pseudomorphic SiGe layer (*i.e.* strained layer below the critical thickness h_C for plastic relaxation) on a SOI substrate. The resulting substrate is then oxidized at high temperature, as in Fig. 2(b). Silicon is oxidized preferentially, while the Ge diffuses through the SiGe layer to the buried oxide. If this process is continued, at some point Si is completely oxidized and the GeOI substrate can be obtained by removing the top SiO_2 layer.

In the case of Ge condensation technique, SOI wafers are only used as substrates for the $Si_{1-x}Ge_x$ layers that are subsequently Ge enriched. Since SOI wafers with thick top Si layers result in more Si atoms in the overall $Si_{1-x}Ge_x/Si$ stack, the Ge enrichment process becomes more time-consuming. This problem can be alleviated by SOI thinning prior $Si_{1-x}Ge_x$ epitaxial growth, with the target for the final top Si thickness being ~20 nm. In order to avoid defect generation, the grown $Si_{1-x}Ge_x$ layer must remain below h_C, which depends on the Ge content x, as shown in Fig. 3.[14] An additional Si cap layer on the $Si_{1-x}Ge_x$ layer can be used to prevent Ge consumption during the oxidation[15] and thereby optimize the Ge enrichment.

To favor Ge diffusion and get continuous homogeneous Ge profiles within the silicon-germanium-on-insulator (SGOI) layers, periodic N_2 annealing steps are added all along the process at 1050 and 900 °C.[16] Figure 3(b) shows high-resolution cross-sectional TEM images of GeOI substrates obtained by Ge enrichment. Neither dislocations, nor stacking faults are seen on the TEM pictures. The Ge film thickness is around 10 nm and may subsequently be increased, *e.g.* by Ge epitaxy, for CMOS processing.[17]

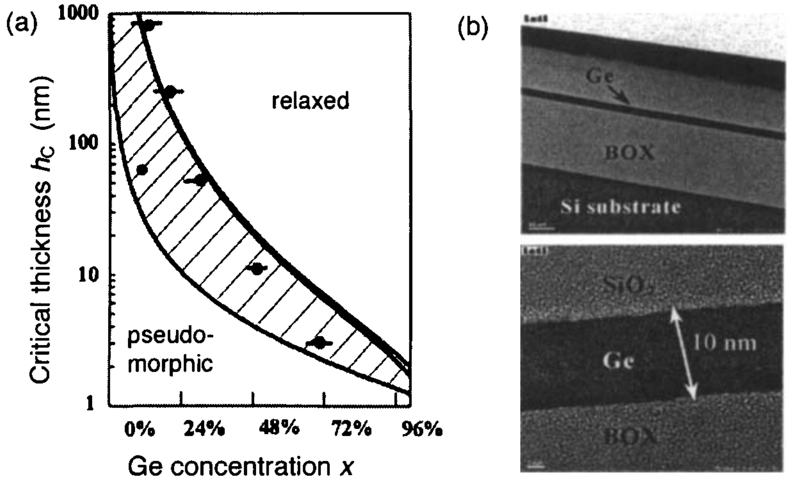


Figure 3. (a) Critical thickness of plastic relaxation of SiGe layer grown epitaxially on Si as a function of the Ge content (dashed region shows metastable films, after Ref. 14); (b) cross-sectional TEM images of 10 nm thick Ge film on the GeOI wafer obtained by Ge enrichment.

3. GeOI material characterization: Ψ-MOSFET

In a world of constant evolution, SOI substrate development needs powerful tools to quickly and accurately characterize the raw material (before CMOS processing). Since a thin Ge film is fully depleted, the Ψ-MOSFET represents the accurate technique to scan the back-interface Ge/buried oxide (BOX) properties and to give crucial data on SOI and GeOI structure quality.[5] The Ge film represents the body of the transistor and the BOX serves as the gate insulator. Two pressure-adjustable metallic probes act as drain and source. A backgate bias V_{BG} induces accumulation or inversion at the BOX-film interface, as shown in Fig. 4(a). The $I_D(V_{BG})$ characteristics permit the extraction of the low-field mobility μ_0, threshold voltage V_T, subthreshold swing, interface state density D_{it}, *etc.* These extractions are carried out using the Y-function method,[18] defined and illustrated in Fig. 4(b). The low-field mobility is calculated from the slope of the linear region of the characteristic, while the intercept of this tangent with the X-axis provides V_T or flat-band voltage V_{FB}. For thin films (< 100 nm), the potential of the front interface (passivated or not) modifies the back-interface potential through the effective electric field present in the film.[19]

The Smart-Cut™ technique is very attractive for BOX engineering and Ge/BOX interface passivation. Indeed, BOX nature can significantly affect the electrical properties like hole and electron mobility, as well as V_T and V_{FB}. Several Ge/BOX interface passivations have been investigated, particularly silicidation and nitridation via inductively-coupled plasma (ICP) treatments.

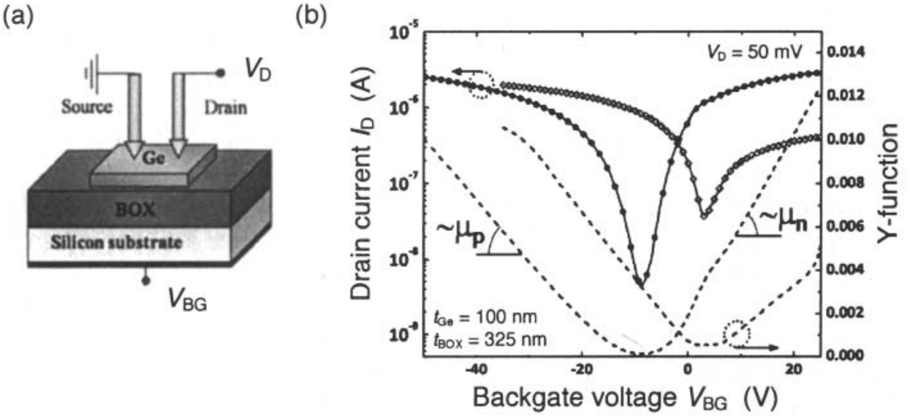


Figure 4. (a) Typical Ψ-MOSFET configuration on GeOI; (b) associated $I_D(V_{BG})$ curves for two GeOI structures. Parameters extraction is performed by the Y-function (= $I_D/g_m^{1/2}$) method.[18]

An example of ICP nitridation leading to a 2 nm GeO_xN_y capping layer at the Ge/BOX interface is shown in Fig. 5(a). Very high electron mobility of ~675 $cm^2/V{\cdot}s$ is achieved, while hole mobility is maintained around 100 $cm^2/V{\cdot}s$.[20] Thus, the GeO_xN_y layer provides high electron currents, giving a possible passivation method for nMOSFETs on Ge or GeOI. The silicon passivation highlights the possibility of 2D Si-capping layer growth on Ge,[21] as well as very good hole mobility improvement at the Ge/BOX interface. Recently, compressively strained Smart-Cut™ GeOI has been demonstrated, improving the hole mobility extracted by Ψ-MOSFET by over 30% (μ_p = 250 $cm^2/V{\cdot}s$ for a 40 nm thick Ge film).[22]

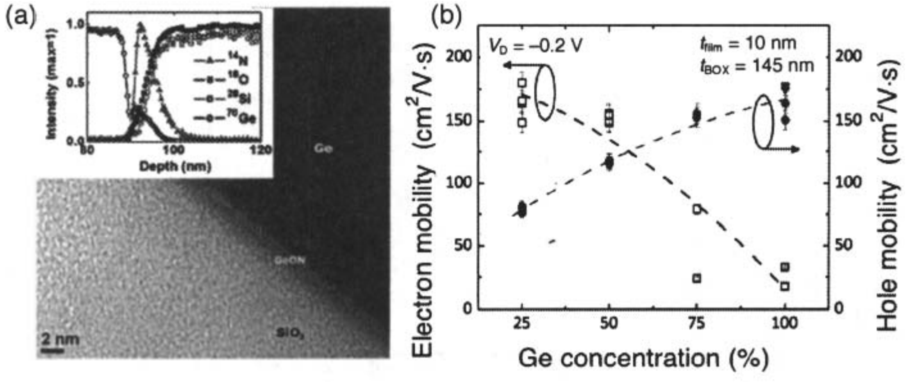


Figure 5. (a) HR-TEM cross-section of the GeO_xN_y capping leading to very high electron and hole back-channel currents (inset shows atomic species found at the back interface);[20] (b) electron/hole low-field mobility extracted by the Y-function on enriched GeOI samples *vs.* Ge content.[23]

The Ge enrichment technique makes it possible to tune the Ge content in the Si_xGe_{1-x} layer, to favor either electron or hole transport properties. Increasing the Ge concentration leads to a diminution of the electron mobility while the hole mobility is enhanced, as shown in Fig. 5(b).[23] In addition, Ge enrichment provides the possibility of straining the epitaxial Ge layer, improving the hole transport properties still further, *e.g.* $\mu_p = 500$ cm²/V·s.

To summarize, both methods of fabricating GeOI substrates have their own advantages and drawbacks. The Smart-Cut™ technique favors the possibility of tuning the back interface with various capping layers, providing high electron mobility substrates, whereas the Ge enrichment technique is more focused on the 2D co-integration and compressively strained substrates. A short overview of the present substrate performances is summarized in Fig. 6,[24] where hole *vs.* electron mobility is plotted to illustrate the results of both methods.

4. High-performance pMOSFETs on GeOI: Achievements and challenges

- *GeOI pMOSFET processing and performance*

Once GeOI substrates are produced as described in the preceding section, pMOS devices can be fabricated using a somewhat modified CMOS process. After Ge stripping for mesa isolation,[25] an additional Ge *n*-type doping via As implantation can be added to prevent front- and back-channel leakage, as well as the threshold voltage shift towards positive gate voltage.[8,26] A ~1 nm Si capping is then deposited on Ge surface prior to the gate stack formation and oxidized to obtain a SiO_x cap over Ge. With the silicon capping approach, it is possible to take advantage of the Si CMOS gate stack, since in this case HfO_2 is deposited on Si/SiO_2.[27] A 4–6 nm thick HfO_2 gate oxide is formed by atomic layer deposition.

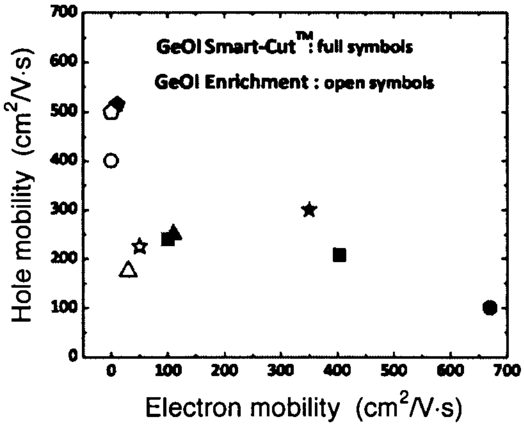


Figure 6. Hole *vs.* electron mobility extracted by Ψ-MOSFET for various GeOI samples (Smart-Cut™ and enriched GeOI). Reproduced from Ref. 24.

The resulting EOT values vary around 2 nm with low gate leakage (I_{OFF} = 0.6 µA/µm at ($V_G - V_T$) = $-V_D/3$).[27] This is followed by 10 nm TiN physical vapor deposition (PVD) and 50 nm Si n^+-poly shunts. After DUV photolithography and gate etching, HfO_2 is removed and BF_2 S/D extensions are implanted and annealed, followed by HDD implantation (dose = 10^{15} cm^{-2}, activated at 600 °C). The HDD implantation step is critical for GeOI pMOS, because *n*-dopants in Ge diffuse quickly[28] and *p*-dopants do not[29] (the converse is true in Si).

We fabricated pMOSFET devices on 200mm GeOI obtained both by Smart-Cut™ (Fig. 7)[30] and by the Ge enrichment technique[8,31] in a fully silicon-compatible process, demonstrating the first L_G = 70 nm FD GeOI pMOS in 2008.[26]

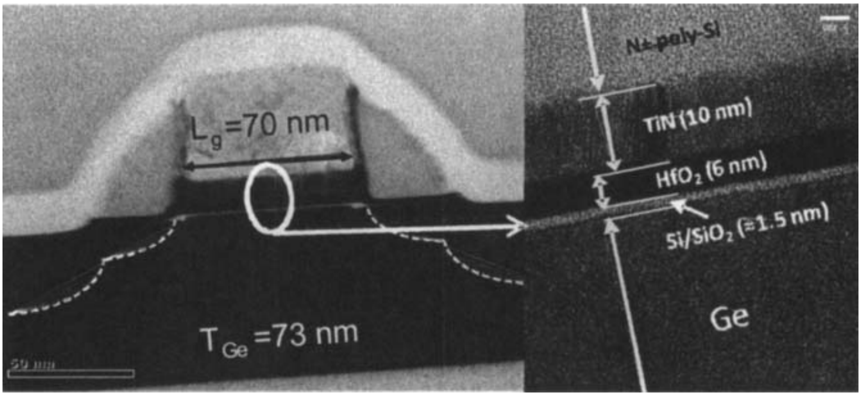


Figure 7. TEM image of 70nm gate length GeOI pMOSFET and HR TEM image of the corresponding $Si/SiO_x/HfO_2/TiN$ gate stack.[26]

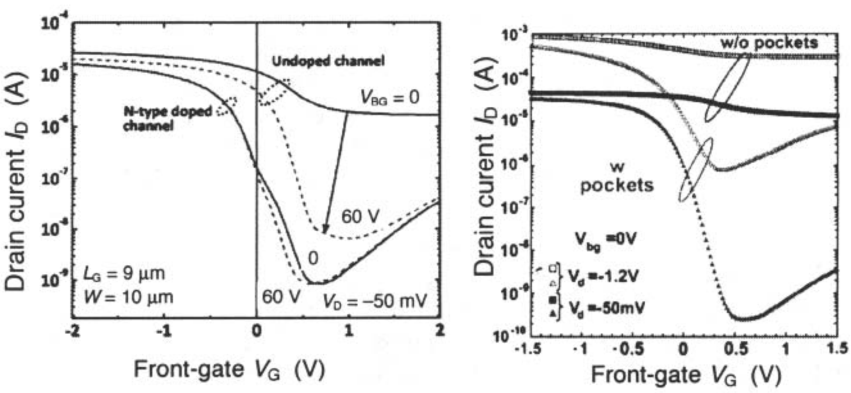


Figure 8. (a) Impact of *n*-type counterdoping of the Ge film on GeOI pMOSFET $I_D(V_G)$ characteristics in linear regime (W = 10 µm, L_G = 9 µm, V_{BG} = 0 or 60 V); (b) impact of pockets on $I_D(V_G)$ characteristics of L_G = 70 nm gate GeOI pMOSFET at V_{BG} = 0 (with *n*-counterdoping).

The electrical characteristics of these transistors are shown in Fig. 8(a). We find that at $V_{BG} = 0$ the back channel is turned on and that I_{OFF} can be significantly reduced by biasing the backgate to $V_{BG} = 60$ V. This effect has been seen in pMOSFETs made on GeOI obtained both by Smart-Cut™[8,30,32] and Ge enrichment.[31,33] The exact origin of this effect is uncertain, but interfacial states at the Ge/BOX interface could be responsible,[7,8,31] as already observed in SOI.[34]

Since the fabrication of 200 mm GeOI wafers is only beginning, the quality of these substrates is not yet comparable to state-of-the-art SOI. As a result, temporary solutions such as channel *n*-type doping can be used to benchmark GeOI CMOS. The efficiency of channel doping is shown in Fig. 8, indicating that the parasitic conduction can be suppressed, making large V_{BG} unnecessary. Likewise, the efficiency of pockets for very aggressive gate length ($L_G = 70$ nm) devices is shown in Fig. 8(b).

For short-channel devices, the drain current is directly impacted by access resistance, which were extracted using Y-function method: $R_{ACC} = R_S + R_D \approx 870$ $\Omega \cdot \mu$m.[26] This value can be reduced by using germanide (*e.g.* NiGe) on S/D areas. By TCAD simulations we determined the expected I_{ON} improvement for the shortest $L_G = 70$ nm device: +28% at $V_D = -1.2$ V with $R_{ACC} = 200$ $\Omega \cdot \mu$m.[26] Also, I_{ON} is affected by the carrier mobility, which we measured as 110 (80) cm^2/V·s for long (short) devices – lower than the 250 cm^2/V·s in undoped GeOI devices, but still better than undoped FD SOI pMOS.[35] Low-temperature measurements, down to $T = 77$ K, show a saturation of the low-field mobility, highlighting a high density of defects along the channel. The temperature dependence of the low-field mobility and I_{ON} are well correlated. Hence, we expect a significant improvement of the mobility, and consequently I_{ON}, once the interface defect problem is solved and channel counterdoping is no longer needed.

The linear I_{OFF} is reduced at low temperature, reflecting the SRH process reduction. As V_D is increased, a high drain leakage appears in the accumulation regime, leading to I_{OFF} current degradation. This leakage is partially suppressed at low T, as expected for a combination of trap-assisted tunneling (lowered at low T) and band-to-band tunneling (BTBT, independent of T).[26] Reducing the trap density is expected to improve I_{OFF} by a decade, but to go further the BTBT component must be minimized, which is a challenge because of Ge's small bandgap. A way to reduce the BTBT is to operate at lower V_D: Fig. 9 shows how I_{OFF} is halved when V_D is reduced from −1.2 to −1 V. More generally, Fig. 9 summarizes the measured and simulated I_{ON}/I_{OFF} trade-offs for Ge and GeOI pMOSFETs.[36,37,38] Simulations show that the addition of a germanide on S/D of short-channel $L_G = 70$ nm pMOSFETs should enable a $I_{ON} = 450$ μA/μm at $I_{OFF} = 100$ nA/μm ($V_D = -1$ V).

Interestingly, a different GeOI approach has recently been reported with ultrathin (3 nm) Ge layer on SOI.[39] This approach counts on quantum confinement in the Ge layer, which increases the effective band gap value and reduces the I_{OFF} current while maintaining a negative V_T without resorting to channel doping.

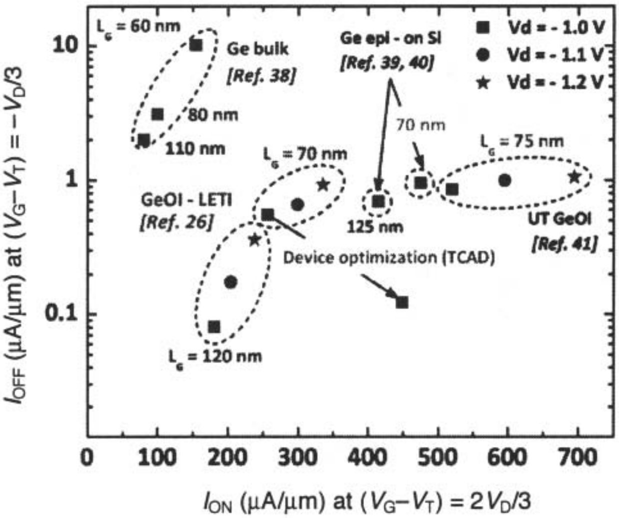


Figure 9. I_{OFF} *vs.* I_{ON} trade-off for GeOI pMOSFETs (L_G = 120 and 70 nm), comparison with bulk Ge reference devices and TCAD simulations.

- *Towards Ge based CMOS?*

In parallel with pMOSFETs, we have fabricated nMOSFETs on GeOI wafers. The nMOSFETs are functional, but the threshold voltage is quite high (+0.8 V), mainly due to large D_{it} at the Ge/gate stack interface. Moreover, the nMOS I_{ON} performance was relatively poor (as generally observed in literature): a factor of 9 smaller than that of pMOS. This disappointing result is due to the reduced electron mobility in the initial substrates: the electron mobility extracted on equivalent GeOI wafers was very low (~10–20 cm²/V·s). This discrepancy between nMOS and pMOS makes straightforward GeOI CMOS integration unviable. Consequently, Ge-based integration schemes generally combine GeOI pFETs with nFETs made in a different semiconductor: silicon or even III-V materials.

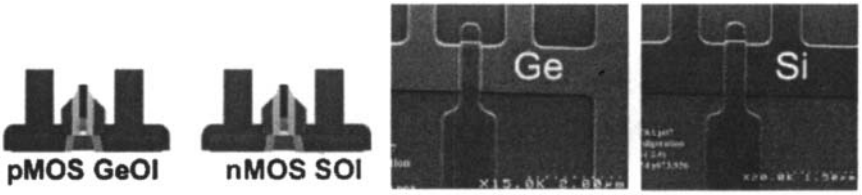


Figure 10. Schematics of the planar co-integration (pFET on GeOI and nFET on SOI on the same level, starting from localized Ge areas on SOI) and corresponding SEM picture of pFET and nFET.[40]

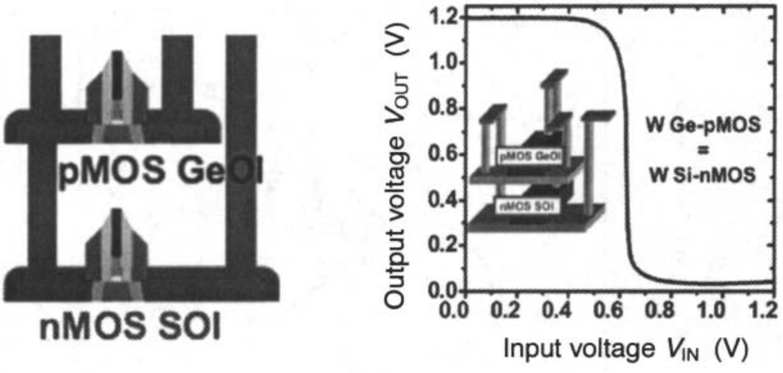


Figure 11. (a) Schematic principle of 3D monolithic co-integration with nMOS on SOI and pMOS on GeOI; (b) inverter transfer characteristic after co-integration.[44]

Figure 10 illustrates one Ge-based CMOS integration scheme involving planar FD-SOI Si nFETs and Ge pFETs. The Ge enrichment technique can create substrates with Si and Ge regions,[40] adapted to this planar co-integration. An optional mobility booster is strain: tensile strain for Si nMOS and compressive strain for Ge pMOS.

Another possibility lies in 3D monolithic integration, illustrated in Fig. 11. Here the nMOSFETs are fabricated on SOI underneath the GeOI layer that is used for pMOSFET fabrication. The process can be summarized as follows. First, nMOSFETs are fabricated on an SOI substrate using a standard high temperature process. After interdielectric deposition and planarization, the Ge film is bonded using a low-temperature molecular process.[41] This 3D co-integration scheme presents the unique advantage of enabling independent optimization of nMOS and pMOS devices. A two-step high-temperature nickel silicidation can be used for nFETs, whereas low temperature germanidation is required for pFETs. The first demonstration, based on a recrystallization technique, was published recently, demonstrating the feasibility of this approach.[42]

This solution is all the more interesting because it potentially leads to a gain in integration density and reduced interconnect delay. An SRAM technology with double-stacked Si layers, shown in Fig. 11(b), has already been demonstrated.[43] Promising 4T-SRAM relying on the coupling between the two stacked layers show enhanced stability and density with respect to other double-gate like architectures.[44] Thus, in conclusion, combining performance gains of GeOI pMOSFETs with monolithic 3D integration advantages promises simultaneous improvements at both the transistor and circuit level.

5. Innovative devices on GeOI: TFETs and nanowires

In order to further lower the power supply voltage V_{DD}, a device based on an alternative operation principle, with a turn-on characteristic sharper than the

60 mV/decade MOS limit, would be very helpful. One such device is the tunneling field-effect transistor (TFET), where the gate voltage controls the interband tunneling between source and drain, and I_{OFF} is not limited by source-drain diffusion. The fundamental idea of the TFET dates back to a proposal by Shockley and Hooper of a gated *pn* junction operated in avalanche mode. Experimentally, TFETs have been realized in III-Vs,[45] bulk Si,[46] SOI,[47] and even carbon nanotubes.[48] Recent simulations indicate that TFETs can have a sharper turn-on characteristic than CMOS transistors,[49] attracting significant interest in the mainstream VLSI community. In SOI, reports of TFETs with $S < 60$ mV/decade have been published,[50,51] with S reaching 42 mV/decade over a limited range of V_G.

The principle of operation of the TFET is explained using the schematic band diagrams in Fig 12. The TFET is an ambipolar device that can operate in two modes depending on the gate bias, with interband tunneling occurring at either the source-channel or the drain-channel junction, although one of the two modes can be suppressed by asymmetric junction engineering. Since both I_{ON} and I_{OFF} are largely determined by the tunneling through an interband barrier, they strongly depend on the maximum electric field E_{MAX} (arising from the built-in junction field combined with the V_D and V_G-induced field components), as well as the channel bandgap E_G.

Due to the relatively large bandgap of Si ($E_G \sim 1.1$ eV), Si-based TFETs, whether SOI or bulk, suffer from relatively low I_{ON}.[50,51] Taking advantage of the narrower Ge bandgap, a TFET in GeOI is in principle expected to exhibit enhanced tunneling I_{ON}. Such devices have been reported recently.[52,53] Figure 13(a) shows the $I_D(V_G)$ characteristics of a GeOI TFET biased at $V_D = 0.8$ V, compared with SOI and SGOI TFETs.[50] The GeOI TFETs show improved I_{ON} currents, albeit with degraded S and I_{OFF}. Figure 13(b) shows $I_D(V_G)$ characteristics of a GeOI TFET with an epitaxial junction (see schematic in figure inset) at various drain biases V_D. In principle, the epitaxial junction should be more abrupt than the standard counterdoped TFET junctions produced by implantation, and the resulting

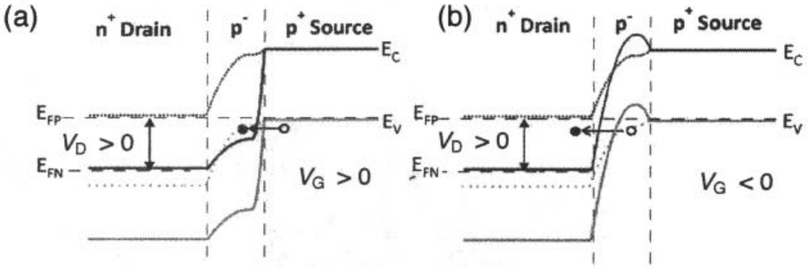


Figure 12. Band diagrams illustrating the interband tunneling in an ambipolar TFET for $V_D > 0$ (reverse bias on the *pn* source-drain junction). (a) For $V_G > 0$, electrons tunnel from the valence band of the p^+-source into the inversion layer of the channel; (b) for $V_G < 0$ electrons tunnel from the accumulation layer of the channel to the conduction band of the n^+-drain. In both cases, dashed lines show the $V_D = V_G = 0$ band alignment.

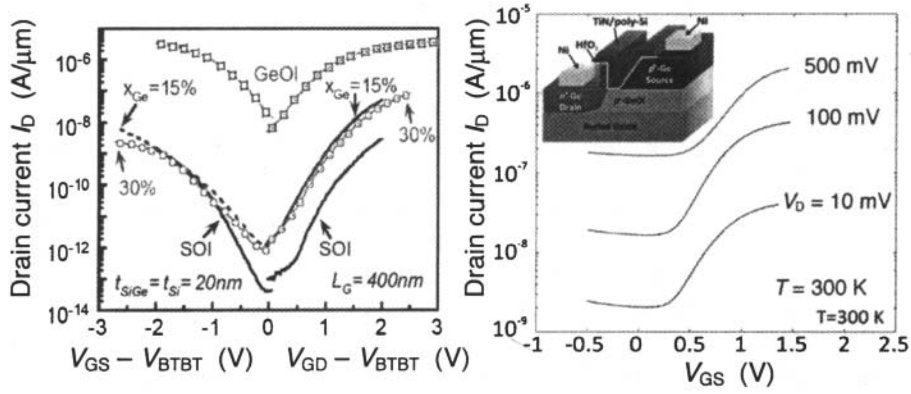


Figure 13. (a) Room-temperature $I_D(V_G)$ characteristics of the GeOI TFET compared to SOI and SGOI TFETs at V_D = 0.8 V;[50] (b) epitaxial GeOI TFET characteristics at several V_D values with the device structure shown in the inset.[52]

I_{ON} is indeed improved, but again at the expense of an inferior S. The same material and interface quality issues that have negatively impacted the performance of GeOI-based CMOS transistors are likely responsible. Note that the architecture enhancements used to improve CMOS performance may likewise benefit TFETs. Thus, Si or Ge nanowires (NWs) offer a promising platform for TFET fabrication. Recently, 3D stacked NW configurations in SOI have been reported with excellent FET characteristics:[54] high I_{ON}, low I_{OFF}, near-ideal S and reduced drain-induced barrier lowering. This type of device layout promises not only superior CMOS, as well as interesting hybrid integration possibilities,[55] but also a platform for high-performance TFET fabrication. In fact, TFETs have been experimentally demonstrated in *in-situ* doped Si NWs grown by vapor-liquid-solid growth,[56] and in scalable vertical Si NWs,[57] showing excellent I_{ON}/I_{OFF} ratios.

Taking advantage of the narrower Ge bandgap, a Ge NW-based TFET is expected to work even better, exhibiting higher current drive. Although Ge NW TFETs have not yet been fabricated, high-quality stacked Ge NWs have been reported[58] using selective etching techniques analogous to those employed for stacked Si NWs – see Fig. 14.

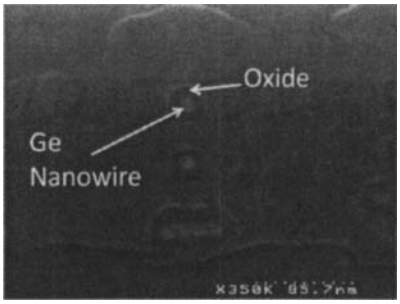


Figure 14. Stacked Ge nanowires (NW) made on GeOI using similar etching techniques to those employed for Si NWs.[58]

6. Conclusions

High-quality GeOI substrates can be produced using both Smart-Cut^{TM} and enrichment techniques. Lower defect densities and Ge/BOX interface passivation are the main benefits of the Smart-Cut^{TM} approach, while enrichment can provide ultrathin and compressively strained Ge layers. Process steps optimization permits the fabrication of high performance 70 nm gate length pMOSFETs on GeOI using high-κ and Si/SiO_x passivation. However several challenges like access resistance, off-state leakage and interface quality remain before GeOI-based devices can reach ITRS specifications for a possible high-power application. To overcome the currently insufficient performance of Ge-based nMOSFETs, Si/Ge co-integration schemes have been developed: vertically using Smart-Cut^{TM} and laterally using Ge enrichment. This co-integration opportunity is a strong advantage for GeOI substrates compared to bulk technologies. Finally, the realization of innovative devices such as Ge nanowires and TFETs can open a path for GeOI in the field of ULSI integration and beyond-CMOS technologies.

Acknowledgments

The work at IMEP and CEA–LETI has been partially carried out within the EUROSOI+ network and Nanosil project ICT-216171. The work at Brown has been supported by the National Science Foundation (ECCS-0701635). One of the authors (A. Z.) gratefully acknowledges support under the RTRA program of the Nanosciences Foundation of Grenoble.

References

1. International Technology Roadmap for Semiconductors (ITRS), 2008 Update Emerging Research Devices (ERD) section.
2. P. Benkart, A. Kaiser, A. Munding, *et al.*, "3D chip stack technology using through-chip interconnections," *IEEE Design Test Comp.* **22**, 512 (2005).
3. H. Shang, M. Frank, E. Gusev, *et al.*, "Germanium channel MOSFETs: Opportunities and challenges," *IBM J. Res. Develop.* **50**, 377 (2006).
4. T. Akatsu, C. Deguet, L. Sanchez, *et al.*, "Germanium-on-insulator (GeOI) substrates: A novel engineered substrate for future high performance devices," *Mater. Sci. Semicond. Processing* **9**, 444 (2006).
5. S. Cristoloveanu and S. Li, *Electrical Characterization of Silicon-on-Insulator Materials and Devices*, Boston: Kluwer Academic Publishers, 1995.
6. J. Mitard, B. De Jaeger, F. E. Leys, *et al.*, "Record I_{ON}/I_{OFF} performance for 65 nm Ge pMOSFET and novel Si passivation scheme for improved EOT scalability," *Tech. Digest IEDM* (2008), art. 4796837.
7. A. Dimoulas and T. Tsipas, "Germanium surfaces and interfaces," *Microelectronic Eng.* **86**, 1577 (2009).

8. W. Van Den Daele, E. Augendre, K. Romanjek, *et al.*, "Low-temperature measurements on germanium-on-insulator pMOSFETs: Evaluation of the background doping level and modeling of the threshold-voltage temperature dependence," *ECS Trans.* **19**, 145 (2009).
9. P. Broqvist, A. Alkauskas, and A. Pasquarello, "Defect levels of dangling bonds in silicon and germanium through hybrid functional," *Phys. Rev. B* **78**, 075203 (2008).
10. R. Xie, T. H. Phung, W. He, *et al.*, "High mobility high-κ/Ge pMOSFETs with 1 nm EOT – New concept on interface engineering and interface characterization," *Tech. Digest IEDM* (2008), art. 4796703.
11. Y. Nakakita, R. Nakane, T. Sasada, H. Matsubara, M. Takenaka, and S. Takagi, "Interface-controlled self-aligned source/drain Ge pMOSFETs using thermally-oxidized GeO_2 interfacial layers," *Tech. Digest IEDM* (2008), art. 4796838.
12. F. Letertre, C. Deguet, C. Richtarch, *et al.*, "Germanium on insulator (GeOI) structures realized by the Smart-Cut^{TM} technology," *Proc. MRS* **809**, 153 (2004).
13. C. Deguet, C. Morales, J. Deschamp, *et al.*, "Germanium on insulator (GeOI) structures realized by the Smart-Cut^{TM} Technology," *Proc. IEEE SOI Conf.* (2004), pp. 96–97.
14. J. C. Bean, L. C. Feldman, A. T. Fiory, S. Nakahara, and I. K. Robinson, "Ge_xSi_{1-x}/Si strained layer supperlattice grown by molecular beam epitaxy," *J. Vac. Sci. Technol. A* **2**, 436 (1984).
15. J. F. Damlencourt and R. Costa, European Patent 0601850.
16. Z. Di, M. Zhang, W. Liu, *et al.*, "Fabrication and mechanism of relaxed SiGe on insulator by modified Ge condensation," *J. Vac. Sci. Technol. B* **23**, 1637 (2005).
17. J. F. Damlencourt, Y. Campidelli, Q. T. Nguyen, *et al.*, "Epitaxial regrowth of Ge on SGOI and GeOI substrates obtained by Ge condensation," *ECS Trans.* **6**, 315 (2007).
18. G. Ghibaudo, "New method for the extraction of MOSFET parameters," *Electronics Lett.* **24**, 543 (1998).
19. G. Hamaïde, F. Allibert, H. Hovel, and S. Cristoloveanu, "Impact of free-surface passivation on silicon-on-insulator buried interface properties by pseudotransistor characterization," *J. Appl. Phys.* **101**, 114513 (2007).
20. T. Signamarcheix, F. Allibert, F. Letertre, T. Chevolleau, L. Sanchez, and E. Augendre, "Germanium oxynitride GeO_xN_y as a back interface layer for GeOI substrates," *Appl. Phys. Lett.* **93**, 022109 (2008).
21. J. M. Hartmann, H. Grampeix, and L. Clavelier, "Epitaxial growth of thick Ge layers on nominal and 6° off Si (001): Ge surface passivation by Si," *ECS Trans.* **16**, 583 (2008).
22. E. Augendre, L. Sanchez, W. Van Den Daele, *et al.*, "Demonstration of compressively-strained GeOI using the Smart-Cut^{TM} technology," submitted to *IEEE SOI Conf.* (2009).

23. Q. T. Nguyen, J. F. Damlencourt, B. Vincent, *et al.*, "High-quality germanium-on-insultor wafers with excellent hole mobility," *Solid State Electronics* **51**, 1172 (2007).
24. E. Augendre, L. Sanchez, L. Benaissa, *et al.*, "Challenges and progress in germanium-on-insulator materials and device development towards ULSI integration," *ECS Meeting Abs.* **902**, 2387 (2009).
25. L. Lachal, J. Chiaroni, C. Le Royer, *et al.*, "Resist stripping process on germanium: A basic post-implant study," *ECS Trans.* **3**, 113 (2006).
26. K. Romanjek, L. Hutin, C. Le Royer, *et al.*, "High performance 70 nm gate length germanium-on-insulator pMOSFET with high-κ/metal gate," *Solid State Electronics* **53**, 723 (2009).
27. C. Le Royer, X. Garros, C. Tabone, *et al.*, "Germanium/HfO$_2$/TiN gate stacks for advanced nodes: Influence of surface preparation on MOS capacitor characteristics," *Proc. ESSDERC* (2005), pp. 97–100.
28. V. Mazzocchi, S. Koffel, C. Le Royer, *et al.*, "Experimental investigation of the impact of implanted phosphorus dose and anneal on dopant diffusion and activation in germanium," *Proc. MRS* **1070**, 35 (2008).
29. L. Hutin, S. Koffel, C. Le Royer, *et al.*, "Activation level in boron-doped thin germanium-on-insulator (GeOI): Extraction method and impact of mobility," *Mater. Sci. Semicond. Proc.* **11**, 267 (2009).
30. C. Le Royer, L. Clavelier, C. Tabone, *et al.*, "0.12 μm p-MOSFETs with high-κ and metal gate fabricated in a Si process line on 200 mm GeOI wafers," *Proc. ESSDERC* (2007), pp. 458–461.
31. W. Van Den Daele, E. Augendre, C. Le Royer, J. F. Damlencourt, B. Grandchamp, and S. Cristoloveanu, "Low temperature characterization and modeling of advanced GeOI pMOSFETs: Mobility mechanisms and origin of the parasitic conduction," to appear in *Solid State Electronics* (2010).
32. C. Le Royer, C. Tabone, K. Romanjek, *et al.*, "105 nm gate length p-MOSFETs with high-κ and metal gate fabricated in a Si process line on 200 mm GeOI wafers," *Solid State Electronics* **52**, 1285 (2008).
33. L. Clavelier, J. F. Damlencourt, C. Le Royer, *et al.*, "High hole mobility GeOI pMOSFETs with high-κ/metal gate on Ge condensation wafers," *Proc. SOI Conf.* (2007), pp. 19-20.
34. J. L. Pelloie, D. K. Sadana, H. J. Hovel, *et al.*, "Effect of the back gate conduction on 0.25 μm SOI devices," *Tech. Digest IEDM* (1994), pp. 653–656.
35. L. Clavelier, C. Deguet, C. Le Royer, *et al.*, "Review of some critical aspects of Ge and GeOI Substrates," *ECS Transactions* **3**, 789 (2006).
36. T. Yamamoto, Y. Yamashita, M. Harada, *et al.*, "High performance 60 nm gate length germanium p-MOSFETs with Ni germanide metal source/drain," *Tech. Digest IEDM* (2007), pp. 1041–1043.
37. G. Nicholas, B. De Jaeger, D. P. Brunco, *et al.*, "High-performance deep submicron Ge pMOSFETs with halo implants," *IEEE Trans. Electron Dev.* **54**, 2503 (2007).

38. G. Hellings, J. Mitard, G. Eneman, *et al.*, "High performance 70-nm germanium pMOSFETs with boron LDD implants," *IEEE Electron Dev. Lett.* **30**, 88 (2009).
39. E. Batail, S. Monfray, C. Tabone, *et al.*, "Localized ultrathin GeOI: An innovative approach to germanium channel MOSFETs on bulk Si substrates," *Tech. Digest IEDM* (2008), pp. 397–400.
40. B. Vincent, J. F. Damlencourt, Q. T. Nguyen, *et al.*, "The Ge condensation technique: A solution for planar SOI/GeOI co-integration for advanced CMOS technologies?" *Mater. Sci. Semicond. Processing* **11**, 205 (2008).
41. J. P. Joply, O. Faynot, and L. Clavelier, "Three-dimensional CMOS integrated circuit and manufacturing method," European Patent EP1811567.
42. J. Feng, Y. Liu, P. B. Griffin, and J. D. Plummer, "Integration of germanium-on-insulator and silicon MOSFETs on a silicon substrate," *IEEE Electron Dev. Lett.* **27**, 911 (2006).
43. S. M. Jung, H. Lim, C. Yeo, *et al.*, "High speed and highly cost effective 72 Mb density S3 SRAM technology with doubly stacked Si layers, peripheral only $CoSi_x$ layers and tungsten shunt W/L scheme for standalone and embedded memory," *Tech. Dig. VLSI Symp.* (2007), p. 82.
44. P. Batude, M. A. Jaud, O. Thomas, *et al.*, "3D CMOS integration: Introduction of dynamic coupling and application to compact and Robust 4T SRAM," *Proc. ICICDT Conf.* (2008), pp. 281–284.
45. T. Uemura and T. Baba, "First observation of negative differential resistance in surface tunnel transistors," *Japan. J. Appl. Phys.* **33**, 207 (1994).
46. W. M. Reddick and G. A. J. Amaratunga, "Silicon surface tunnel transistor," *Appl. Phys. Lett.* **67**, 494 (1995).
47. C. Aydin, A. Zaslavsky, S. Luryi, *et al.*, "Lateral inteband tunneling transistor on silicon-on-insulator," *Appl. Phys. Lett.* **84**, 1780 (2004).
48. J. Appenzeller, J. Knoch, M. Radosavljević, and Ph. Avouris, "Multimode transport in Schottky-barrier carbon-nanotube field-effect transistors," *Phys. Rev. Lett.* **92**, 226802 (2004).
49. Q. Zhang, W. Zhao, and A. Seabaugh, "Low-subthreshold swing tunnel transistor," *IEEE Elecron Dev. Lett.* **27**, 297 (2006).
50. W. Y. Choi, B.-G. Park, J. D. Lee, and T.-J. King Liu, " Tunneling field-effect transistors (TFETs) with subthreshold swing (SS) less than 60 mV/dec," *IEEE Electron Dev. Lett.* **28**, 743 (2007).
51. F. Mayer, C. Le Royer, J. F. Damlencourt, *et al.*, "Impact of SOI, $Si_{1-x}Ge_xOI$ and GeOI substrates on CMOS compatible tunnel FET performance," *Tech. Digest IEDM* (2008), pp. 163–166.
52. T. Krishnamohan, D. Kim, S. Raghunathan, and K. Saraswat, "Double-gate strained-Ge heterostructure tunneling FET (TFET) with record high drive currents and <60 mV/dec subthreshold slope," *Tech. Digest IEDM* (2008), art. 4796839.

53. D. Kazazis, P. Jannaty, A. Zaslavsky, *et al.*, "Tunneling field-effect transistor with epitaxial junction in thin germanium-on-insulator," *Appl. Phys. Lett.* **94**, 263508 (2009).
54. C. Dupré, A. Hubert, S. Bécu, *et al.*, "15 nm-diameter 3D stacked nanowires with independent gate operation: ΦFET," *IEDM Tech. Digest* (2008), art. 4796805.
55. T. Ernst, L. Duraffourg, C. Dupré, *et al.*, "Novel Si-based nanowire devices: Will they serve ultimate MOSFETs scaling or ultimate hybrid integration?" *IEDM Tech. Digest* (2008), art. 4796804.
56. M. T. Björk, J. Knoch, H. Schmid, H. Riel, and W. Riess, "Silicon nanowire tunneling field effect transistors," *Appl. Phys. Lett.* **92**, 193504 (2008).
57. Z. X. Chen, H. Y. Yu, N. Singh, *et al.*, "Demonstration of tunneling FETs based on highly scalable vertical silicon nanowires," *IEEE Electron Dev. Lett.* **30**, 754 (2009).
58. E. Saracco, J.-F. Damlencourt, D. Lafond, *et al.*, "Fabrication of suspended Ge-rich nanowires by Ge condensation technique for multi-channel devices," *ECS Trans.* **19**, 207 (2009).

This page intentionally left blank

Simulation of Self-Heating Effects in Different SOI MOS Architectures

Enrico Sangiorgi, Marco Braccioli, and Claudio Fiegna
ARCES-DEIS University of Bologna and *IUNET*
Via Venezia 260, 47521 Cesena, Italy

1. Introduction

The demand for high circuit performance has been addressed historically by miniaturization of device dimensions and increased device packing densities.[1,2]

The International Technology Roadmap of Semiconductors (ITRS)[3] imposes aggressive scaling trends for the gate length (L_G) and equivalent gate oxide thickness (EOT) to achieve a 17% annual reduction of the intrinsic switching delay (CV/I) for high-performance transistors. Over time, this strategy has become insufficient because of mobility degradation due to the increasing channel-doping density. Furthermore, it has become increasingly difficult to control undesirable phenomena, such as leakage current across the gate oxide, short-channel effects causing an excessive I_{OFF} current (the subthreshold current in a transistor that is nominally off), a large dependence of threshold voltage V_T on drain bias, and the variability of V_T due to statistical fluctuations of physical parameters like channel length and the number of dopant atoms in the channel.[4]

Technology scaling also impacts power dissipation. The classic constant-field scaling rule predicts a significant reduction of the dynamic power in a given circuit, proportional to fCV_{DD}^2, whereas the power dissipation per unit area is expected to stay constant. Actual scaling trends have relaxed the reduction of supply voltage in order to comply with the ITRS requirements in terms of gate delay, therefore leading to a continuous increase of dynamic power density.[5]

The power crisis has become even worse due to increased subthreshold drain and gate leakage currents, such that the static power density has become comparable with the dynamic power contribution.[5-7]

This increase in power dissipation translates into larger heat generation and rising temperatures. Heat dissipation typically occurs through the silicon bulk toward the backside of the die, and thermal energy is ultimately removed by a heat sink.

The relevance of thermal effects in device design increases with the introduction of new MOSFET device architectures. In fact, in order to overcome the increasing difficulties in achieving improvements in terms of device and circuit performance, substantial innovations are currently proposed for the near future, including new materials to be adopted in the fabrication of both the wafer and the single device, and modifications to the device architecture. In the short term,

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

adoption of new materials for the realization of the gate stack and the channel of conventional bulk and partially depleted (PD)-SOI MOSFETs is suggested: high-κ dielectrics allow EOT scaling while keeping the tunneling gate leakage under control,[8] whereas new channel materials can improve mobility,[9,10] thus increasing the I_{ON}/I_{OFF} ratio. Additional innovations in device architecture are under scrutiny, with the intent to introduce ultrathin-body (UTB) fully-depleted (FD) SOI devices and multiple gate devices, where the short-channel effects (SCE) are effectively suppressed by scaling down the thickness of the silicon layer.[11,12]

All these innovations must be carefully analyzed from the point of view of heating effects. The adoption of SiGe alloys or thick buried-oxide (BOX) layers in SOI wafers significantly impacts the thermal properties of the device, as these materials possess a much lower thermal conductivity than bulk silicon. In addition, it has been pointed out that ultrathin silicon layers with thickness t_{Si} on the order of a few tens of nanometers, required in order to cope with the SCE in FD single-gate (SG) and double-gate (DG) devices, present a degraded thermal conductivity due to phonon boundary scattering that reduces the phonon mean free path significantly.[13-16]

This paper discusses self-heating effects in different SOI architectures by 3D electrothermal simulations and compares different device architectures, such as planar SG and DG transistors, as well as fin-FETs.

2. Heat generation in short devices

Heat generation in MOSFETs occurs through the emission of phonons by carriers heated by the electric field in the channel region. Since the electric field is largest at the drain end of the channel and the source-to-drain distance is comparable to the mean free path for phonon scattering (~10 nm), each carrier suffers few if any scattering events while travelling across the intrinsic channel region (quasi-ballistic transport). Hence, most of the heat generation takes place inside the drain. The most common approach for the analysis of heat generation consists of calculating the heating rate per unit volume as the scalar product of the electric field and current-density vectors ($J \cdot F$). This approach is strictly local in its nature and fails to take into account the nonlocal characteristics of carrier heating and phonon emission. The local model predicts that the maximum heat-generation rate takes place at the electric field peak, that is, at the drain junction. More accurate calculations of phonon emission performed by Monte Carlo (MC) transport simulations predict a much broader heat-generation region, with a lower peak value, displaced inside the drain junction.[17] Nonlocal effects on carrier heating and phonon emission become more relevant as the device L_G approaches the mean free path for phonon emission. The precise quantification of the impact of the local-heating approximation is still an open issue because of the lack of a detailed treatment of short- and long-range electron–electron interaction that influences the carrier heating and the subsequent cooling process (thermalization) occurring in the drain region. In spite of its inherently local approximation, the conventional model for Joule heating is still widely used in device simulators, as it allows for efficient electrothermal simulation via the self-consistent coupling of carrier transport, heat generation, and heat transport. This makes it possible to investigate the impact of technological options on device performance, including the self-heating effects (SHE).[18] More work is needed in order to quantify the effects of nonlocal heat generation and to include them in the electrothermal device simulation.

3. Transport of heat generated in nanoscale semiconductor devices

The transport of heat in semiconductors is due to the propagation of lattice vibrations (phonons), whereas the contribution by electrons (dominant in the case of metals) is estimated on the order of 1% even in the case of very large concentrations.[19] Different temperatures at two positions in a semiconductor device imply different distributions of phonons. Since the change in phonon distribution may only occur due to scattering, the temperature may vary only over a length larger than the phonon mean free path. The simple phenomenological approach based on the Fourier's law of heat diffusion is the most widely adopted one for modeling the heat transport,

$$C_S \partial T/\partial t = \nabla(k_S \nabla T) + H(\boldsymbol{r}, t)\,, \tag{1}$$

where C_S and k_S are the heat capacity per unit volume and the thermal conductivity of the semiconductor, respectively, and H represents the heat-generation rate per unit volume. The thermal conductivity can be written as

$$k_S = C_S v \Lambda_S/3\,, \tag{2}$$

where Λ_S represents the phonon mean free path (approximately 200–300 nm in undoped bulk silicon at 300 K), and v is the average phonon velocity.

When studying the heat transport in nanoscale devices with dimensions on the order of, or smaller than, the phonon mean free path, sub-continuum transport effects occur[20] as follows:

- Hot-spot ballistic phonon-emission effects: heat generation due to optical phonon emission by heated electrons, taking place in a region of limited extension (hot spot) compared with the phonon mean free path, leads to a higher temperature rise compared to predictions of the Fourier diffusion theory because a significant change in temperature may occur only over a distance comparable to or larger than the phonon mean free path.
- Hot spot far from equilibrium effects: SHE occurs mainly through the emission of optical phonons characterized by a low propagation velocity. On the other hand, the heat transport involves much faster acoustic phonon modes. The large rate of the optical phonon emission at the drain end of the channel and the difference in the propagation velocity of the optical and acoustic phonons lead to a localized nonequilibrium condition, in which the ratio of optical-phonon concentration to acoustic-phonon

concentration is much larger than at equilibrium. This results in a reduced heat-spreading capability.

- Phonon boundary scattering in thin films: thermal conductivity in thin films is substantially reduced with respect to bulk crystals due to the enhanced scattering of phonons at film boundaries, causing a large reduction of the phonon mean free path.

The first two issues challenge the continuum diffusion theory of heat transport represented by Eq. (1). A higher order treatment of heat transport, which is able to cope with the aforementioned hot-spot-related issues, would require the solution of several phonon Boltzmann transport equations (one for each phonon mode) coupled with each other by the phonon scattering. This approach is hindered both by the complexity of the Boltzmann transport equation (BTE) when applied to realistic structures, and the limited knowledge about the selection rules and transition rates for phonon-phonon interactions. Several simplified approaches for the simulation of heat transport based on the phonon BTE have been proposed, and most of them are discussed in Ref. 21. In particular, the split-flux model[22] has been applied to the analysis of heat transport in short MOSFETs.[23]

The enhanced boundary scattering in thin films leads to a reduced phonon mean free path, thus reducing the effects of hot-spot ballistic phonon emission and making the limitations of Eq. (1) less critical. The enhanced scattering can be taken into account in the frame of the simple diffusion theory (1) by appropriately modifying the thermal conductivity in order to account for the enhanced boundary scattering. Given the experimental data of thermal conductivity in the ultrathin silicon layers[15,16] we will now discuss a physics-based model for the effect of thickness-dependent boundary scattering on the thermal conductivity.

4. Simulation approach

Three-dimensional (3D) electrothermal (ET) simulations have been performed using the Sentaurus device simulator.[24] The model proposed by Canali and co-workers[25] is employed to describe high-field transport; mobility degradation at the Si/SiO$_2$ interface due to surface roughness scattering, and mobility dependence on doping are included as well. The drift-diffusion (DD) model underestimates I_{ON} in nanometer MOSFETs operating in the quasi-ballistic regime, due to increasing importance of nonequilibrium transport phenomena that occur when the gate length is scaled down. For this reason, the mobility model parameters have been calibrated in order to reproduce MC-calculated $I_{DS}(V_{DS})$ characteristics, following the approach suggested by Bude.[26] In particular, the saturation velocity and the exponent that describes the temperature dependence of the saturation velocity in the conventional model for high-field-dependent carrier mobility have been calibrated in order to achieve a good agreement between DD and MC results. The calibration of the DD transport model has been performed at different temperatures by comparison with a validated, full-band self-consistent MC simulator for a 3D electron gas.[27,28] We followed the approach employed in Ref. 29 in order to calibrate the high-field-dependent mobility model for FD SG SOI transistors featuring 25 and 18 nm L_G. All the remaining mobility model parameters related to the dependence on doping concentration and surface roughness are kept at their default values.

The 3D ET simulations are computationally intensive. In order to perform simulations that satisfy the requirements in terms of CPU time and allocated memory, quantum corrections (even if available through a density gradient approach) are not taken into account. Moreover, in order to minimize the node count of the structures, the simulation domain is only one-half of the complete device (*i.e.* we simulated only $W_{fin}/2$ in fin-FETs and only $W_{ch}/2$ in planar devices). This is possible by exploiting the symmetries of the simulated structures.

In our ET simulations, the silicon thermal conductivity k_{Si} depends on film thickness and temperature; in particular, k_{Si} is a decreasing function of these two quantities. Liu[16] has presented experimental data of thermal conductivity in ultra-thin silicon layers. In the same work, a model for the effect of thickness-dependent boundary scattering on the thermal conductivity is proposed and validated for a silicon thickness down to 20 nm. By adopting this model, we extrapolated k_{Si} values for silicon thicknesses down to 5 nm and for temperatures ranging between 250 and 500 K. As for the thermal boundary conditions adopted in our simulations, an isothermal (IT) 300 K heat sink is placed at the bottom of the underlying 1.8 mm thick Si layer. The gate and S/D contacts are treated either as adiabatic or 300 K IT boundary conditions; moreover, a lumped thermal resistance $R_{TH} = 2\times10^{-4}$ cm^{2}K/W is connected between the gate and a 300 K IT boundary condition. This lumped resistance takes into account the thermal resistance due to the gate dielectric and to the gate/SiO_2 interface and it does not depend on the gate insulator thickness.[30] Finally, the vertical x-y and z-y planes (see Fig. 1) are treated as adiabatic boundaries.

5. Comparative analysis of SHE in different SOI architectures

In this section n-channel SG-SOI, DG-SOI and fin-FETs are compared in terms of self-heating effects by calculating $I(V)$ curves by both isothermal and electro-thermal simulations. Figure 1 provides a simple sketch of the simulated planar devices and fin-FETs; Table 1 summarizes the values assumed for the most important device parameters.[31]

The simulation domain is 14 mm wide in the direction along the channel; we have performed simulations featuring 8 and 20 mm wide domains as well, and we have found that the differences in terms of obtained current and temperature are less than 1%. This trend is common to all the considered structures, and shows that the simulation domain adopted does not impact the results. A 1.8 mm thick Si substrate layer is placed beneath the BOX. An abrupt junction is assumed between the S/D regions and the substrate.

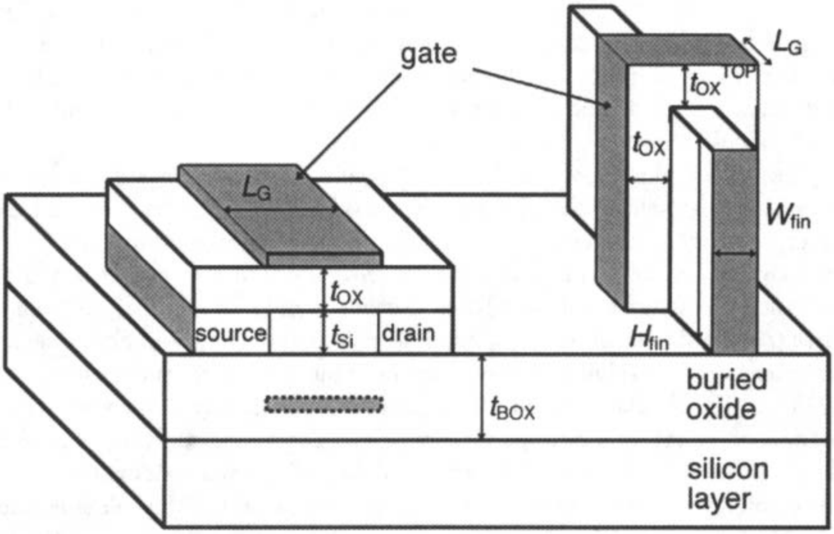


Figure 1. Three-dimensional sketch of the simulated SG SOI, DG SOI and fin-FET (not to scale). The grey-shadowed dashed line region represents the second gate for the DG SOI transistor.

	DG SOI	SG SOI	fin-FET
Gate length L_G (nm)	50	64/50	50
Gate workfunction ϕ_G (eV)	4.6	4.62/4.61	4.6
Gate oxide t_{OX} (nm)	1	1/1	1
Gate oxide dielectric constant ε_{ox}	3.9	3.9/3.9	3.9
Silicon thickness t_{Si} (nm)	10	10/5	n.a.
Channel width W_{ch} (nm)	250	250/250	n.a.
Fin width W_{fin} (nm)	N/A	N/A	10
Fin height H_{fin} (nm)	N/A	N/A	60
Top-gate thick. t_{OX}^{TOP} (nm)	N/A	N/A	5
S/D access length (nm)	35	35	35
S/D doping (cm^{-3})	10^{20}	10^{20}	10^{20}
Channel doping (cm^{-3})	10^{15}	10^{15}	10^{15}
Buried oxide t_{BOX} (nm)	50	50	50
Supply voltage V_{DD} (V)	1.0	1.0/1.0	1.0

Table 1. Key parameters adopted in the comparison presented in Section 4. Third column reports two different SG SOI architectures, defined to have the same isothermal electrical characteristics as DG SOI and fin-FET.

As we are interested in the impact of self-heating, care has been taken in order to ensure the same IT electrical characteristics for all the devices, that is the same threshold voltage V_T, the same transconductance and a good tolerance to the SCE (*i.e.* we have designed the structures in order to have drain-induced barrier lowering values lower than 100 mV/V).

A pure SiO_2 gate dielectric with $t_{OX} = 1$ nm has been chosen. In the fin-FET, the top gate features an oxide much thicker than the lateral ones ($t_{OX}^{TOP} = 5$ nm), so that the contribution to the drain current due to the top gate is negligible. The DG SOI and fin-FET transistors feature the same $L_G = 50$ nm, as well as the same silicon thickness ($t_{Si} = W_{fin} = 10$ nm).

The SGSOI architecture is strongly affected by SCE. In order to overcome this problem, two different solutions in designing the SG SOI transistor have been adopted: the first one (SG-1) features $t_{Si} = 10$ nm but a slightly longer $L_G = 64$ nm than the DG SOI and fin-FET; the second one (SG-2) features the same L_G as DG SOI and fin-FET but a much thinner silicon body, $t_{Si} = 5$ nm. While the fin–FETs' fin width is set to 10 nm, the planar SOI devices feature a channel width $W_{ch} = 5L_G$, a realistic minimum width for planar devices.

Figure 2 shows the transfer and output characteristics, obtained from IT simulations. The devices feature almost the same IT currents per unit width, as required for a consistent comparison. Conversely, Figure 3 shows the transfer and output characteristics obtained from ET simulations. In these simulations, the gate and S/D contacts are taken as adiabatic, so the power generated in the active region can be dissipated only through the BOX. The SHE-related degradation of I_{DS} strongly depends on the device structure: it is maximum for the DG SOI, less critical for the fin-FET, and minimum for SG-1. This trend is confirmed by Fig. 4, where the maximum temperature rise $\Delta T_{MAX} \equiv (T_{MAX} - 300\text{ K})$, with T_{MAX} being the maximum temperature reached inside the device, is plotted as a function of the dissipated power P per unit width. The slope of each line can be interpreted

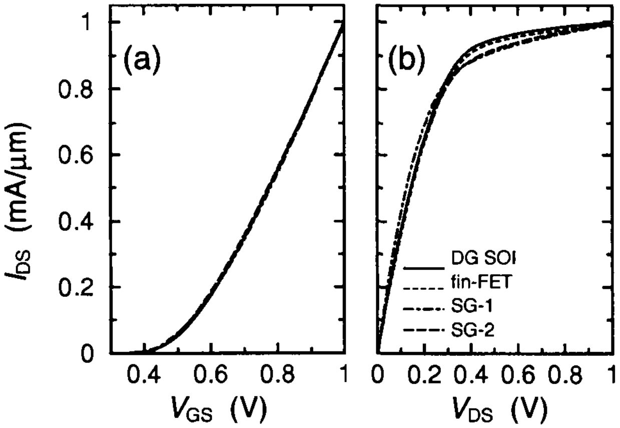


Figure 2. (a) Transfer characteristics at $V_{DS} = 1.0$ V and (b) output characteristics at $V_{GS} = 1.0$ V, calculated by 3D drift-diffusion isothermal simulations. The figure confirms that the different devices feature similar $I(V)$ curves.

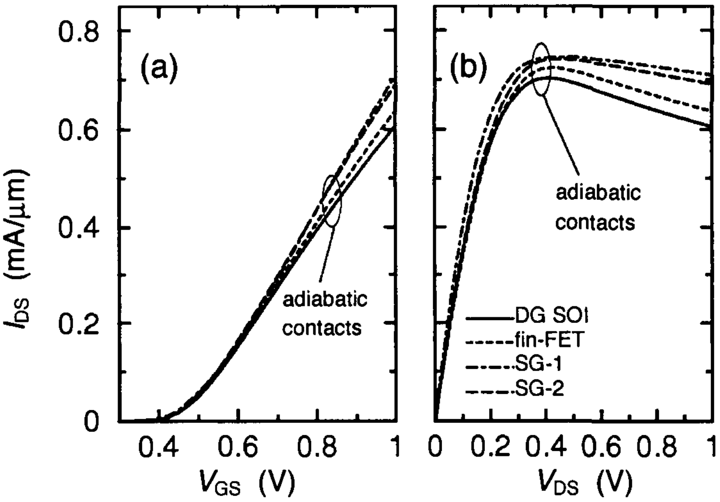


Figure 3. (a) Transfer characteristics at $V_{DS} = 1.0$ V and (b) output characteristics at $V_{GS} = 1.0$ V, calculated by 3D drift-diffusion electrothermal simulations. The gate and S/D contacts are treated as adiabatic and the heat flux can occur only through the buried oxide.

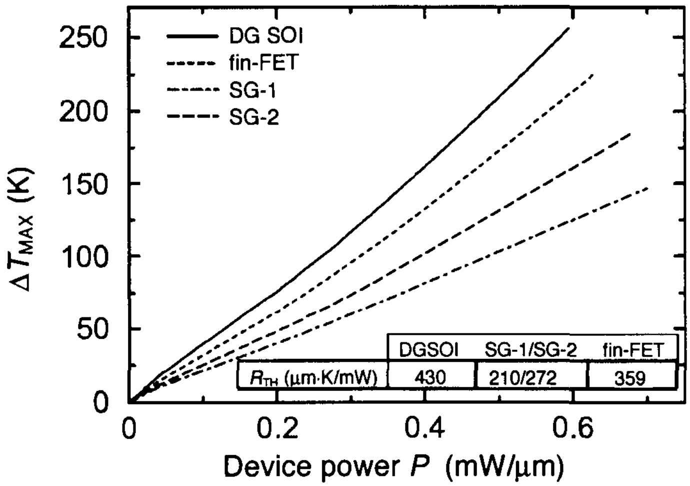


Figure 4. $\Delta T_{MAX} = (T_{MAX} - 300\text{ K})$ *vs.* dissipated power per unit width P, for the four considered devices; the gate and S/D contacts are treated as adiabatic. As in Fig. 3, SG-1 presents the lowest temperature rise and equivalent thermal resistance R_{TH} (*i.e.* $R_{TH} = \Delta T_{MAX}/P$). The inset in the figure shows R_{TH} values for $V_{GS} = V_{DS} = V_{DD}$.

as the thermal resistance R_{TH} associated to the corresponding device; the table inset in the same figure reports R_{TH} values for $V_{GS} = V_{DS} = V_{DD}$. The DG SOI presents the highest thermal resistance, SG-1 the lowest one. The differences between SG SOI and the other two structures are related to the wider overlap area between the Si body and the underlying BOX, through which most of the heat flux occurs. This area is smaller in DG SOI and fin-FET (in particular, it is halved in DG SOI, with respect to SG SOI), and this degrades the heat dissipation. Furthermore, the larger SHE of SG-2 compared to SG-1 is due to the degradation of k_{Si} occurring as t_{Si} is scaled down from 10 nm in SG-1 to 5 nm in SG-2. The fin-FET presents lower SHE than DG SOI because W_{fin} is much lower than W_{ch}, see Table 1. In this case, heat dissipation in the direction of device width is larger in the fin-FET than in the DG SOI. The relevance of this component increases as the device width decreases, as can be seen in Fig. 5: as W_{ch} of the DG SOI is scaled down, the values of I_{ON} and T_{MAX} depart from those obtained from the 2D DGSOI simulation (corresponding to infinite W_{ch}), and approach the fin-FET values.

Figure 6 shows the transfer and output characteristics, obtained from ET simulations, when the S/D and gate contacts are treated as 300 K boundary conditions. In this case they act as cooling contacts, and they contribute to dissipating the heat generated in the active region. The differences between the devices in terms of drain current degradation become less evident compared to Fig. 3: most of the cooling occurs through the S/D contacts and the cross-sectional area of the S/D access regions become the most relevant parameter, leading to larger SHE in the SG-2 and DG SOI cases, compared to SG-1 and fin-FET. For the forthcoming technological nodes, the distance between the active region and the S/D contacts is expected to be scaled down; this trend could exploit the cooling occurring through the contacts and contribute to keep SHE under control.

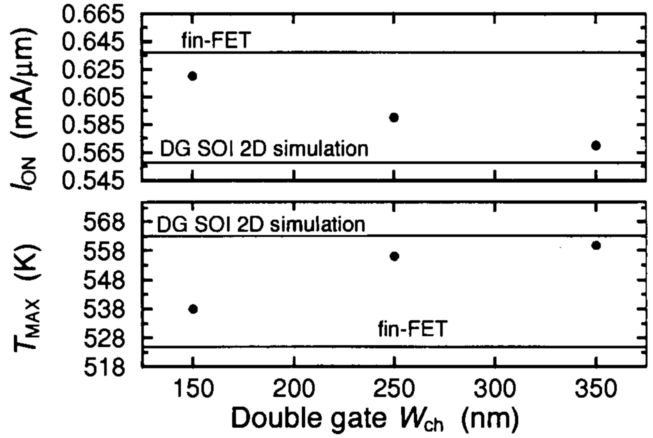


Figure 5. I_{ON} at $V_{GS} = V_{DS} = V_{DD}$ (top) and T_{MAX} (bottom) *vs.* DG SOI channel width W_{ch}. The gate and S/D boundary conditions are assumed adiabatic, so the heat dissipation can occur only through the BOX. As W_{ch} decreases, the values depart from the 2D DG SOI simulation case (infinite W_{ch}) and approach the fin-FET case.

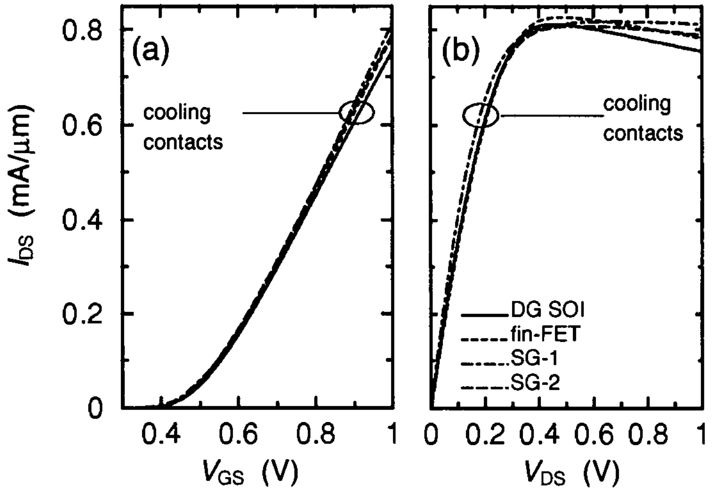


Figure 6. (a) Transfer characteristics at V_{DS} = 1.0 V and (b) output characteristics at V_{GS} = 1.0 V, calculated by 3D DD electrothermal simulations. The gate and S/D contacts are treated as 300 K isothermal boundary conditions.

6. Conclusions

We performed 3D fully-coupled electrothermal simulations of different SOI structures. Planar SG and DG as well as fin-FET transistors have been analyzed, in order to explore how self-heating effects impact the device performance of these different architectures. The results have been compared to reference isothermal electrical characteristics. Our simulations show that SHE detrimentally impacts the device performance and that it is structure-dependent. In particular, the thermal resistance associated with the S/D access regions, as well as the ratio between the surface available for the vertical heat dissipation through the buried oxide and the volume of the active region, largely differ from structure to structure.

References

1. R. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circ.* **9**, 256 (1974).
2. G. Baccarani, M. R. Wordeman, and R. Dennard, "Generalized scaling theory and its application to a 1/4 micrometer MOSFET design," *IEEE Trans. Electron Dev.* **31**, 452 (1984).
3. The International Technology Roadmap for Semiconductors, 2005 ed., is available at *http://www.itrs.net*
4. A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Trans. Electron Dev.* **50**, 1837 (2003).
5. E. J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM J. Res. Develop.* **46**, 169 (2002).
6. V. De and S. Borkar, "Technology and design challenges for low power and high performance," *Proc. ISLPED* (1999), pp. 163–8.
7. W. Haensch, E. J. Nowak, R. H. Dennard, *et al.*, "Silicon CMOS devices beyond scaling," *IBM J. Res. Develop.* **50**, 339 (2006).
8. E. P. Gusev, V. Narayanan, and M. M. Frank, "Advanced high-κ dielectric stacks with poly-Si and metal gates: Recent progress and current challenges," *IBM J. Res. Develop.* **50**, 387 (2006).
9. H.-N. Lin, H.-W. Chen, C.-H. Ko, C.-H. Ge, H.-C. Lin, T.-Y. Huang, and W.-C. Lee, "Channel backscattering characteristics of strained PMOSFETs with embedded SiGe source/drain," *Tech. Digest IEDM* (2005), pp. 141–4.
10. S. E. Thompson, M. Armstrong, C. Auth, *et al.*, "A logic nanotechnology featuring strained-silicon," *IEEE Electron Dev. Lett.* **25**, 191 (2004).
11. R.-H. Yan, A. Ourmazd, and K. F. Lee, "Scaling the Si MOSFET: From bulk to SOI to bulk, " *IEEE Trans. Electron Dev.* **39**, 1704 (1992).
12. K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling theory for double-gate SOI MOSFETs, " *IEEE Trans. Electron Dev.* **40**, 2326 (1993).
13. M. Asheghi, M. N. Touzelbaev, K. E. Goodson, Y. K. Leung, and S. S. Wong, "Temperature-dependent thermal conductivity of single-crystal silicon layers in SOI substrates," *J. Heat Transfer* **120**, 30 (1998).
14. Y. S. Ju and K. E. Goodson, "Phonon scattering in silicon films with thickness of order 100 nm," *Appl. Phys. Lett.* 74, 3005 (1999).
15. W. Liu and M. Asheghi, "Thermal conductivity measurements of ultrathin single crystal silicon layers," *J. Heat Transf*er **128**, 75 (2006).
16. W. Liu, K. Etessam-Yazdani, R. Hussin, and M. Asheghi, "Modeling and data for thermal conductivity of ultrathin single-crystal SOI layers at high temperature," *IEEE Trans. Electron Dev.* **53**, 1868 (2006).
17. E. Pop, J. Rowlette, R. W. Dutton, and K. E. Goodson, "Joule heating under quasi-ballistic transport conditions in bulk and strained silicon devices," *Proc. SISPAD* (2005), pp. 307–10.

18. S. Kumar, R. V. Joshi, C.-T. Chuang, K. Kim, J. Y. Murthy, K. T. Schonenberg, and E. J. Nowak, "Self-consistent and efficient electrothermal analysis for poly/metal gate fin-FETs," *Tech. Digest IEDM* (2006), pp. 803–6.
19. G. K. Wachutka, "Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling," *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.* **9**, 1141 (1990).
20. D. G. Cahill, W. K. Ford, K. E. Goodson, *et al.*, "Nanoscale thermal transport," *J. Appl. Phys.* **93**, 793 (2003).
21. S. Sinha and K. E. Goodson, "Review: Multiscale thermal modeling in nanoelectronics," *Intern. J. Multiscale Comput. Eng.* **3**, 107 (2005).
22. S. Sinha, E. Pop, and K. E. Goodson, "A split-flux model for phonon transport near hotspots," *ASME Electronic Photonic Packaging* **4**, 75 (2004).
23. J. Rowlette, E. Pop, S. Sinha, M. Panzer, and K. Goodson, "Thermal simulation techniques for nanoscale transistors," *Proc. ICCAD* (2005), pp. 225–8.
24. Sentaurus Synopsis, ver. X-2005.10.
25. C. Canali, G. Majni, R. Minder and G. Ottaviani, "Electron and hole drift velocity measurements in silicon and their empirical relation to electric field and temperature," *IEEE Trans. Electron Dev.* **22**, 1045 (1975).
26. J. D. Bude, "MOSFET modeling into the ballistic regime," *Proc. SISPAD* (2000), pp. 23–6.
27. P. Palestri, D. Esseni, S. Eminente, C. Fiegna, E. Sangiorgi, and L. Selmi, "Understanding quasi-ballistic transport in nano-MOSFETs: Part I – scattering in the channel and in the drain," *IEEE Trans. Electron Dev.* **52**, 2727 (2005).
28. S. Eminente, D. Esseni, P. Palestri, C. Fiegna, L. Selmi, and E. Sangiorgi, "Understanding quasi-ballistic transport in nano-MOSFETs: Part II – technology scaling along the ITRS," *IEEE Trans. Electron Dev.* **52**, 2736 (2005).
29. C. Fiegna, Y. Yang, E. Sangiorgi, and A. G. O'Neill, "Analysis of self-heating effects in ultrathin-body SOI MOSFETs by device simulation," *IEEE Trans. Electron Dev.* **55**, 223 (2008).
30. P. He, L. Liu, and Z. Li, "Measurement of thermal conductivity of buried oxides of silicon-on-insulator wafers fabricated by separation by implantation of oxygen technology," *Appl. Phys. Lett.* **81**, 1896 (2002).
31. M. Braccioli, G. Curatol, Y. Yang, E. Sangiorgi, and C. Fiegna, "Simulation of self-heating effects in different SOI MOS architectures," *Solid State Electronics* **53**, 445 (2009).

Nanowires: Technology, Physics and Perspectives

D. Grützmacher, Th. Schäpers, S. Mantl, S. Feste, Q. T. Zhao, H. Hardtdegen, R. Calarco, M. Lepsa, and N. Demarina

Institute for Semiconductor Nanoelectronics and *Jülich Aachen Research Alliance*
Forschungszentrum Jülich, D-52425 Jülich, Germany

1. Introduction

Nanowires (NWs) are currently considered as strong contenders to provide possible solutions in several areas of semiconductor research, including the integration of III-V compounds on Si, the ultimate scaling of CMOS devices, high speed and low power circuitry ("green transistor"), and the realization of quantum information technology. This versatility of NWs has attracted the attention of numerous research groups studying their fabrication and characterization. However, even though substantial progress has been made in recent years, including the demonstration of many device prototypes, some of the fundamental NW physics is not yet fully understood in areas as varied as NW nucleation, structural, transport and optical properties.

Silicon NWs are grown typically either by the vapor-liquid-solid (VLS) mechanism using Au particles as catalyst[1-3] or are fabricated by nanolithography.[4] The latter approach makes it possible to carve NWs into strained Si,[5] leading to uniaxially strained structures. The NW can be wrapped with a surrounding gate, providing optimized control of the transistor channel due to improved electrostatics compared to planar CMOS technology. Further, the NW technology may enable us to integrate novel transistor concepts, like Schottky barrier transistors or tunnel transistors,[6,7] to optimize carrier injection, speed and power consumption. Nanopatterned Si substrates have great potential in serving as compliant templates for the growth of III-V nanostructures. The nanoscale footprint provided by these templates can accommodate a large lattice mismatch by elastic deformation.

Indium-rich III/V nanowires, namely In(Ga)N, In(Ga)As and In(Ga)Sb NWs, are of particular interest due to their high electron mobility and low effective mass. The latter leads to the observation of quantum effects in reasonably "large" NWs.[8] These In-rich NWs have the peculiarity that the Fermi level at the surface is pinned within the conduction band, leading to metallic behavior in the conductivity for narrow wires, with an electron gas at the circumference. Indeed, magnetotransport investigations reveal pronounced oscillations in the conductivity periodic in magnetic flux. These are explained by the contribution of coherent circular quantum states along the surface of the nanowires. However, little is known about

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

the shape of the potential in these In-rich nanowires and its dependence on the doping level and the diameter of the nanowire.

Finally, the large surface to volume ratio in NWs leads to a strong impact of surface scattering on the mobility. Thus, the mobility reported in In-rich nanowires falls short compared to quantum well or even bulk structures. Possible solutions are core/shell structures providing a well-defined clean interface for electron confinement. First experiments performed using MOVPE-grown GaAs/AlGaAs core/shell structures will be discussed. The introduction of lateral and vertical heterointerfaces while growing the nanowires are envisioned as a first step to exactly control the growth and positioning of quantum dots. Ultimately, this route may provide a suitable concept for quantum information processing.

2. Silicon nanowires

We have fabricated silicon NW devices using a top-down approach. A fairly standard process has been employed, using e-beam lithography to define the wire, followed by reactive ion etching (RIE) for the nanowire mesa. Next the gate stack is formed using either a SiO_2/poly-Si gate stack or a high-κ dielectric film combined with a matching metallic layer. Finally, source and drain contacts are implanted and metallized. Figure 1 shows a scanning electron microscopy (SEM) image of a long channel device fabricated using a strained Si-on-insulator (SSOI) layer. Long-channel devices were fabricated to accurately determine the impact of strain on the mobility of charge carriers in the channel. To estimate the strain in the Si nanowire carved into an SSOI layer, finite element simulations of the strain in suspended Si NWs have been performed. It is obvious that the strain perpendicular to the NW axis (*x*-direction in Fig. 2 (left)) is released due to plastic deformation, whereas the strain parallel to the axis (*y*-direction in Fig. 2 (left)) is expected to remain, *i.e.* the wire is stretched between the source and drain contacts, leading to a uniaxial strain within the Si wire.

The contour plots show the results of the finite element calculations for wires with different aspects ratios, *i.e.* wire height *vs.* wire width in *x*-direction, ranging from 8 to 32.[5] Figure 2 (right) shows that for the *x*-component a small compressive strain is found, whereas strong tensile strain persists along the Si wire.

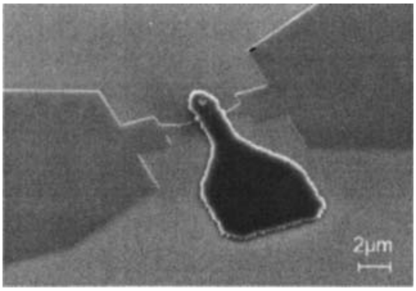


Figure 1. SEM image of a Si nanowire FET with a top gate.

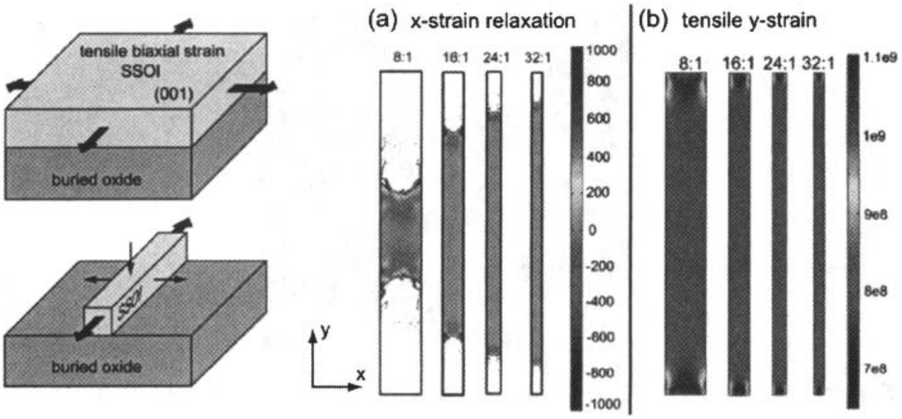


Figure 2. Illustration of the strain relaxation for Si nanowires fabricated from strained SOI layers.

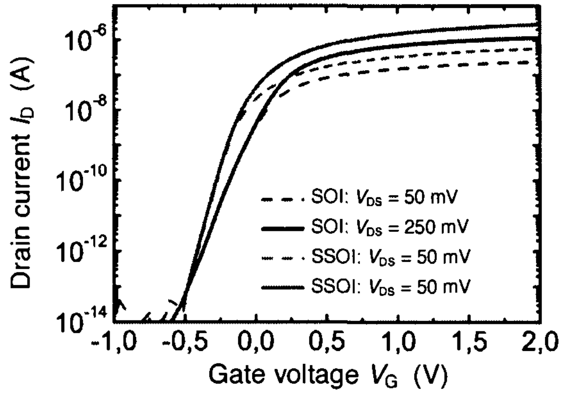


Figure 3. Comparison of $I(V_G)$ characteristics at fixed V_{DS} of Si NW devices fabricated from strained and unstrained SOI layers (L = 3 μm for all devices).

The calculations indicate that the strong uniaxial strain along the Si NW axis should lead to reduced electron effective mass and thus enhanced electron mobility. To study the impact of the strain, devices have been fabricated from both SSOI and unstrained SOI substrates. Figure 3 shows the $I(V)$ characteristics of both types of devices processed using the same parameters.

The devices have a 3 μm long channel and are measured under a source-drain voltage V_{DS} of 50 (dashed lines) and 250 mV (solid lines). For both conditions, the saturation current increases by a factor of 2.5 for strained devices compared to unstrained devices. This result clearly indicates that the incorporated strain in the Si channel is substantially larger than that observed for mechanically bent Si NW transistors[9] and exceeds 1 GPa as indicated in the model calculation presented in Fig. 2 (right).

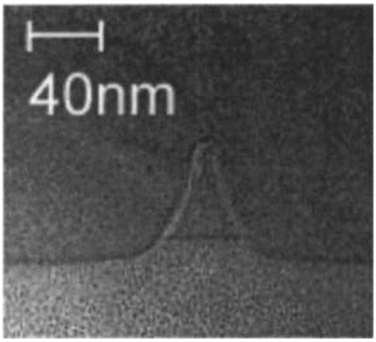


Figure 4. Cross sectional TEM image of a Si nanowire fabricated on SSOI.

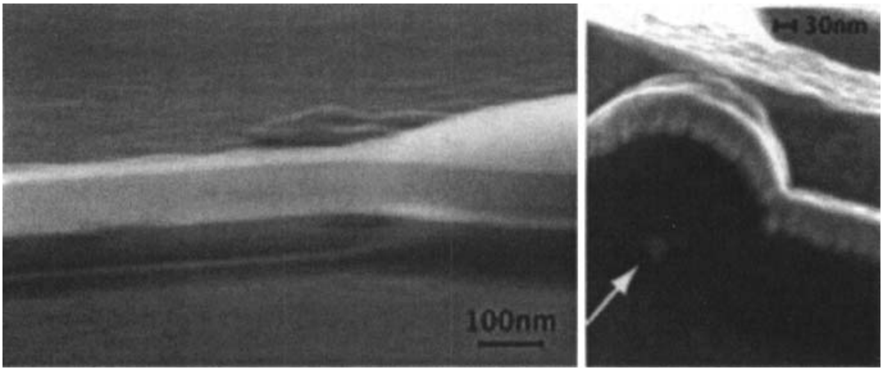


Figure 5. SEM images of freestanding Si nanowire (left) and circular nanowire after several oxidation and etching cycles (right).

Short-channel devices were fabricated from 50 nm thick SSOI layers. Devices having a gate length L_G = 0.2 μm show a nearly ideal subthreshold slope of 62 mV/dec and excellent I_{ON}/I_{OFF} ratios exceeding 10^7 at V_{DS} = 250 mV. Figure 4 shows a cross-sectional TEM micrograph of the channel, which has a triangular shape with the gate wrapping the sidewalls of the device. Further improvements can be expected for devices with an all-around gate. For this purpose, wires with a rectangular cross-section have been freed from the substrate and transformed into a wire with a circular cross-section, as shown in Fig. 5, using repeated oxidation and etching steps. The optimized electrostatics of a circular device with an all-around gate should lead to an even better I_{ON}/I_{OFF} ratio.

The developed technology is suitable for the further integration of new device concepts and designs, in particular Schottky-MOSFETs[6] and tunnel FETs.[7] These concepts may allow modified carrier injection mechanisms that can overcome the 60 mV/dec limit for the subthreshold slope in standard FETs. Moreover, the quantization of the states in a nanowire might be beneficial for the increase of the saturation currents in a tunnel FET (TFET).

3. InN nanowires

Nitride nanowires can easily be grown by MBE on a Si substrate using appropriate growth conditions. Typically N-rich conditions are used to grow GaN[10] as well as InN[11] NWs. The presence of nitrogen radicals released by the plasma cell leads to a formation of an amorphous SiN_x layer at the Si surface. Despite this amorphous layer, crystalline nanowires with the c-axis perpendicular to the substrate surface nucleate on Si (111). Moreover, for GaN NWs even a thick amorphous SiO_2 layer does not hamper the wire formation.[12] Here we discuss the InN NWs deposited on Si (111), which exhibit a surface accumulation layer. The Fermi level is pinned about 0.8–0.9 eV above the conduction band edge,[13] leading to a pronounced band bending and a cylindrical accumulation layer containing an electron gas along the wire surface. The situation is depicted in Fig. 6. The conduction band edge with the corresponding electron wavefunctions are calculated by solving the coupled Poisson and Schrödinger equations in cylindrical coordinates using the effective mass approximation.

The calculations shown in Fig. 6 correspond to an InN nanowire with a diameter of 80 nm (radius $r = 40$ nm) and a surface state density of 10^{13} cm^2. The Fermi level in the centre of the wire is slightly below the conduction band edge. These conditions lead to the filling of two subbands E_1 and E_2, with all carriers confined in the vicinity of the surface. The situation can be considered as a two-dimensional electron gas (2DEG) scrolled into a narrow tube. Due to strong band bending, the carrier confinement at the nanowire surface is very robust and present even in wires having a diameter less than 50 nm. The cylindrical 2DEG was analyzed by magnetotransport measurements with the magnetic field B both normal and parallel to the NW axis. For a B field parallel to the axis, the energy levels for the electrons are periodic in the magnetic field. The periodicity is given

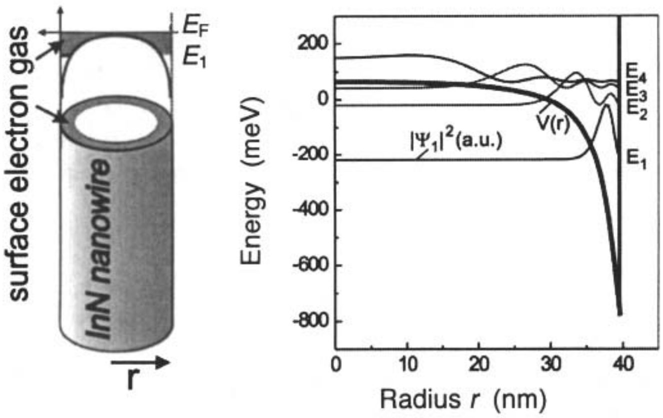


Figure 6. Schematic view of the surface enhancement layer in an InN NW and calculated profile of the conduction band $V(r)$ and probability density for the lowest four subbands $n = 1$–4. The Fermi level is set to the zero of energy.

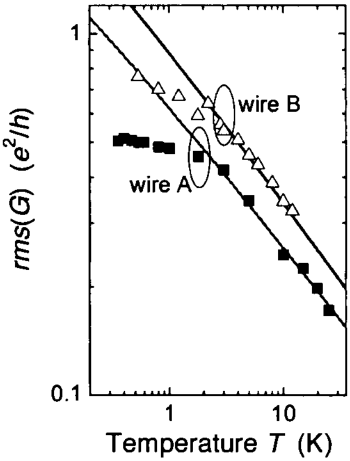


Figure 7. Root mean square of the universal magnetoconductance *G* oscillations *vs.* temperature *T*. For $T < 2$ K the amplitude saturates as the phase coherence length exceeds the wire length *L*.

by the magnetic flux quantum $\Phi_0 = h/e$ and consequently oscillations of the conductivity in dependence on magnetic flux[14] are expected. Well-resolved oscillations in the resistivity as a function of B are indeed detected, with the periodicity of Φ_0/A, where A denotes the area of the cross-section of the wire. These oscillations persist up to a temperature of 30 K. The obtained data clearly indicate that the electrons move in a tubular 2DEG at the vicinity of the nanowire surface. At low bias voltages in the diffusive transport regime, the electrons walk randomly in the 2DEG, having a free mean path of l_C between scattering events as well as a phase coherence length $l_\Phi > l_C$ defined by phase breaking scattering events, similar to planar 2DEG structures.[15] Both quantities strongly depend on temperature due to phonon scattering. However, in nanostructures the dimensions L of the sample might be close to l_Φ, leading to universal conductance fluctuations (UCFs) depending on a perpendicular magnetic field.[16] These UCFs are most pronounced when the sample size and l_Φ are similar. Accordingly, the $rms(G)$ roughness of the UCF provides insight into the phase coherence length of the electrons in the diffusive regime:

$$rms(G) = \alpha \frac{e^2}{h} \left(\frac{l_\Phi(T)}{L} \right)^{3/2} .$$

In Fig. 7, the temperature dependence of the *rms* roughness is plotted for two InN wires of different length L. The increase of the *rms* value as the temperature drops reflects the increase of l_Φ with decreasing T. For both wires a saturation of the rms(G) value is found at low temperatures; at this point the l_Φ value exceeds the length of the wire.[17]

The magnetotransport measurements reveal coherent quantum states in the InN nanowires, which might be attributed to the nearly perfect crystallinity and low surface defect density of the nanowires prepared by self-assembled MBE. Moreover, the low-temperature quantum transport properties of gated InN nanowires were investigated. A clear weak antilocalization effect[18] was found in the averaged magnetoconductance indicating the presence of spin-orbit coupling.[19] The latter effect is important for spin manipulation in nanoscale spintronic devices. The unexpected robustness of the weak antilocalization effect observed in narrow InN wires gives evidence that the tubular topology of the surface electron gas in InN nanowires is maintained also at very small dimensions.

4. InAs nanowires

Indium arsenide NWs were deposited by MBE on GaAs (111) wafers covered with very thin (~6 nm) layers of HSQ, a resist that forms a SiO_x layer upon annealing. Such thin layers will exhibit a large density of pinholes, which act as nucleation points for NWs.[20] It is speculated that in the first stage of growth, an In droplet forms in the depression of the pinhole. The wire starts to grow via the VLS growth mode, with the In droplet remaining on top of the growing wire. Figure 8 (left) shows an SEM image of the sample after deposition of InAs wires. Some multifaceted InAs clusters are found on the surface, next to the wires. It is unclear whether they grew on the SiO_x surface or at pinholes of a size not suitable for wire growth. The InAs wires have a length of a few micros and diameters in the ~100 nm range. Astonishingly, they are rather uniform in size, which might support the assumption that only pinholes of suitable size lead to wire growth. The InAs have a pronounced faceting exhibiting [110] sidewalls, which can be clearly seen in the enlargement of a single wire in Fig. 8 (right).

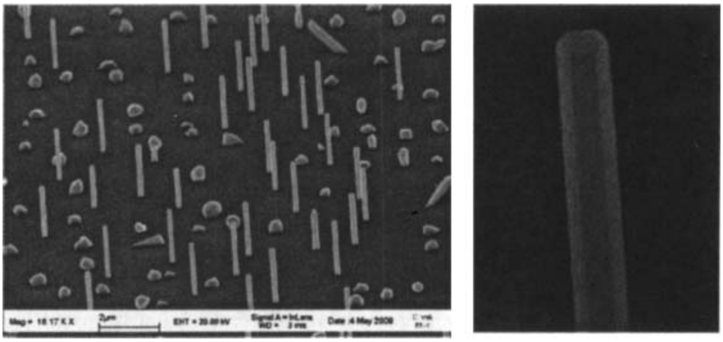


Figure 8. SEM images of MBE grown InAs nanowires on GaAs (111) substrates covered with a thin SiO_x layer.

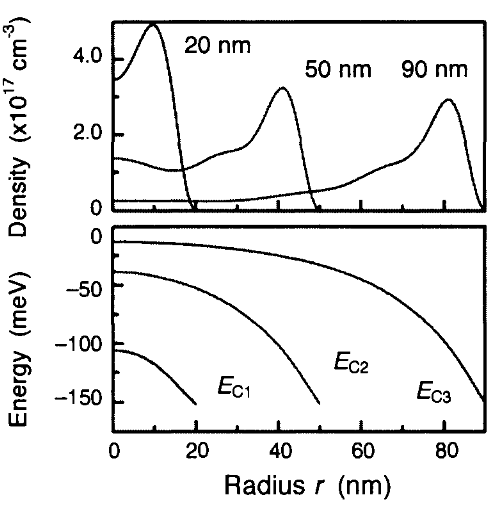


Figure 9. Calculated profile of the conduction band (lower curves) for three InAs NWs with 20, 50 and 90 nm radius and corresponding total electron densities (upper curves) as a function of radius.

In contrast to InN wires, the surface state density of InAs is lower and the Fermi level at an InAs surface is pinned only about 150 meV above the conduction band edge. Again a surface enrichment layer is formed leading to a tubular 2DEG at the surface of an undoped InAs nanowire. However, this 2DEG is much less confined and hence loses its clear 2D character at small diameters as compared to InN wires. The smaller pinning energy and the smaller electron mass lead to larger spreading of the electron wave function across the wire. As a result, the electron gas is transformed from a tubular 2DEG configuration to a quasi 1D electron gas at a wire thickness below 50 nm.

This is illustrated in Fig. 9. The lower graph shows the conduction band edges E_{C1}, E_{C2}, and E_{C3} across the wire for three different wire radii $r = 20$, 50 and 90 nm, respectively. The Fermi energy is located at 0 meV in this illustration. The conduction band edge for the $r = 20$ nm is located more than 100 meV below the Fermi energy even in the centre of the wire, consequently the electron density (upper graph in Fig. 9) in the centre of the wire amounts to 3.5×10^{17} cm^{-3} for an undoped wire. Even though the maximum electron concentration in this wire is still tubular, a quasi one-dimensional transport can be expected. For larger wire diameters the tubular shape becomes more pronounced, however, even for a 180 nm wide nanowire the conduction band edge stays below the Fermi energy at the centre of the wire, making the entire wire degenerate. Consequently, undoped InAs nanowires are expected to have metallic properties, in particular the conductivity should increase with decreasing temperature. Experimentally it is found that the conductivity decreases, *i.e.* the resistance increases by approximately 30% with decreasing temperature, thus opposite to the predicted metallic

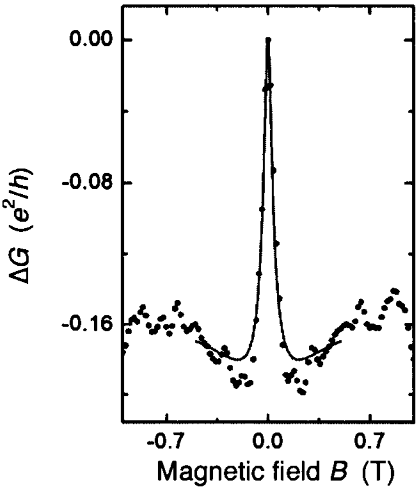


Figure 10. Averaged magnetoconductance G at low magnetic fields. Solid line represents the fit obtained by using Kettemann's model.[22]

behavior. Most likely, the carrier mobility of the electrons in these InAs wires is dominated by temperature-independent surface scattering. The increase in resistance of thin wires might be attributed to a decrease in carrier density at low temperatures caused by an activation energy dependence of electrons released by surface states. However, further studies are required to clarify this phenomenon. It should be noted, that Si-doped InAs wires show a pronounced reduction of their resistivity compared to undoped InAs wires. Furthermore the doped wires exhibit the expected metallic behavior, *i.e.* an increase of the resistance with temperature.

Magnetotransport measurements of thin InAs wires exhibit no signature of weak antilocalization. We attribute this finding to the predominantly one-dimensional transport properties within these wires. For InAs nanowires with a larger diameter, where the majority of the carriers is located in the tubular 2DEG configuration, quantum interference effects have been observed, similar to the results described above for the InN nanowires. The effective electric field in the tubular 2DEG due to the band bending leads to spin precession. Figure 10 shows the results of magnetotransport studies depicting the averaged conductivity G as a function of small B perpendicular to the wire axis. A clear peak at $B = 0$ is found, indicating weak antilocalization, *i.e.* spin precession leads to net destructive interference of backscattered electrons in a 2DEG.[21]

5. Conclusion

In this chapter we discussed the potential for semiconductor quantum wires for future nanoelectronics. Si NWs are a possible path to scale CMOS technology to its ultimate limits. The integration of novel device concepts, like Schottky barrier transistors or TFETs, modifying the carrier injection mechanisms to achieve ultra fast device switching frequencies are promising. However, to achieve high performance 1D transport properties are crucial. Thus technological challenges to achieve Si wires with less than 7 nm diameter and controlled doping profiles, as well as appropriate source/drain and gate configurations, are quite serious.

Several paths for the fabrication of self-assembled III-V NWs grown without the use of a catalyst have been demonstrated. The possibility of growing such NWs on both III-V and Si substrates may allow future integration of opto-electronics on Si as well as advanced beyond-CMOS devices exploring spin as fundamental carrier of logic information.

References

1. A. M. Morales, C. M. Lieber, "A laser ablation method for the synthesis of crystalline semiconductor nanowires," *Science* **279**, 208 (1998).
2. T. Martensson, M. Borgstrom, W. Seifert, B. J. Ohlsson, and L. Samuelson, "Fabrication of individually seeded nanowire arrays by vapor-liquid-solid growth," *Nanotechnology* **14**, 1255 (2003).
3. H. J. Fan, P. Werner, and M. Zacharias, "Semiconductor nanowires: From self-organization to patterned growth," *Small* **2**, 700 (2006).
4. H. Sellier, G. P. Lansbergen, J. Caro, *et al.*, "Subthreshold channels at the edges of nanoscale triple-gate silicon transistors," *Appl. Phys. Lett.* **90**, 073502 (2007).
5. S. F. Feste, J. Knoch, D. Buca, and S. Mantl, "Fabrication of uniaxially strained silicon nanowires," *Thin Solid Films* **517**, 320 (2008).
6. M. Zhang, J. Knoch, J. Appenzeller, and S. Mantl, "Improved carrier injection in ultrathin-body SOI Schottky-barrier MOSFETs," *IEEE Electron Dev. Lett.* **28**, 223 (2008).
7. S. Saurabh and M. J. Kumar, "Impact of strain on drain current and threshold voltage of nanoscale double gate tunnel field effect transistor: Theoretical investigation and analysis," *Japan. J. Appl. Phys.* **48**, 064503 (2009).
8. A. Fuhrer, C. Fasth, and L. Samuelson, "Single electron pumping in InAs nanowire double quantum dots," *Appl. Phys. Lett.* **91**, 052109 (2007).
9. Y. J. Jeong, K. Miyaji, T. Saraya, and T. Hiramoto, "Silicon nanowire n-type metal-oxide-semiconductor field-effect transistors and single-electron transistors at room temperature under uniaxial tensile strain," *J. Appl. Phys.* **105**, 084514 (2009).
10. R. K. Debnath, R. Meijers, T. Richter, T. Stoica, R. Calarco, and H. Lüth, "Mechanism of molecular beam epitaxy growth of GaN nanowires on Si(111)," *Appl. Phys. Lett.* **90**, 123117 (2007).
11. R. Calarco and M. Marso, "GaN and InN nanowires grown by MBE: A comparison," *Appl. Phys. A* **87**, 499 (2007).

12. T. Stoica, E. Sutter R. J. Meijers, R. K. Debnath, R. Calarco, H. Lüth, and D. Grützmacher, "Interface and wetting layer effect on the catalyst-free nucleation and growth of GaN nanowires," *Small* **4**, 751 (2008).
13. V. N. Brudnyi, A. V. Kosobutsky, and N. G. Kolin, "The charge neutrality level and the Fermi level pinning in A(3)N (BN, AlN, GaN, InN) nitrides," *Russ. Phys. J.* **51**, 1270 (2008).
14. Th. Richter, Ch. Blömers, H. Lüth, R. Calarco, M. Indlekofer, and Th. Schäpers, "Flux quantization effects in InN nanowires," *Nano Lett.* **8**, 2834 (2008).
15. S. Washburn and R. A. Webb, "Quantum transport in small disordered samples from the diffusive to the ballistic regime," *Rep. Prog. Phys.* **55**, 1311 (1992).
16. Y. Ochiai, T. Onishi, M. Kawabe, K. Ishibashi, J. P. Bird, Y. Aoyagi, and T. Sugano, "Interference area of universal conduction fluctuations in narrow GaAs/AlGaAs wires," *Japan. J. Appl. Phys.* **32**, 528 (1993).
17. Ch. Blömers, Th. Schäpers, T. Richter, R. Calarco, H. Lueth, and M. Marso, "Temperature dependence of the phase-coherence length in InN nanowires," *Appl. Phys. Lett.* **92**, 132102 (2008).
18. G. Petersen, S. Estévez Hernández, R. Calarco, N. Demarina, and Th. Schäpers, "Spin-orbit coupling and phase-coherent transport in InN nanowires," *Phys. Rev. B* **80**, 125321 (2009).
19. Z. W. Jia, W. Z. Shen, H. Ogawa, and Q. X. Guo, "Electron dephasing in wurtzite indium nitride thin films," *Appl. Phys. Lett.* **89**, 232107 (2006).
20. S. Yoshida, I. Tamai, T. Sato, and H. Hasegawa, "Selective molecular beam epitaxy growth of GaAs hexagonal nanowire networks on (111)B patterned substrates," *Japan. J. Appl. Phys.* **43**, 2064 (2004).
21. S. Estevez Hernandez, M. Akabori, K. Sladek, *et al.*, "Spin-orbit coupling and phase-coherence in short InAs nanowires," submitted to *Phys. Rev. B* (2009).
22. S. Kettemann, "Dimensional control of antilocalization and spin relaxation in quantum wires," *Phys. Rev. Lett.* **98**, 176808 (2007).

This page intentionally left blank

Emerging Nanotechnology for Integration of Nanostructures in Nanoelectronic Devices

T. Baron, C. Agraffeil, F. Dhalluin, M. Kogelschtaz, G. Cunge, T. Chevolleau, B. Salem, B. Salhi, H. Abed, A. Potié, L. Latu-Romain, C. Ternon, K. Aissou
Labo. des Technologies de la Microélectronique, CNRS & Univ. Joseph Fourier
17 Rue des Martyrs, F38054 Grenoble, France

L. Montès
IMEP-LAHC/Minatec, Grenoble-INP, 3 parvis Louis Neel, Grenoble, France

P. Mur, G. Molas, B. De Salvo, E. Jalaguier, T. Ernst, P. Ferret
CEA–LETI/Minatec, 17 Rue des Martyrs, F38054 Grenoble, France

P. Gentile and N. Pauc
CEA-INAC/SiNaPS, 17 Rue des Martyrs, F38054 Grenoble, France

1. Introduction

Scaling down of semiconductor devices is the driving force for the development of new applications (mobile phone, memory cards, sensors, *etc.*). Up to now, device downscaling has been enabled by the top-down approach, combining lithography and etching technologies. However, we will soon reach the limitation of this approach to form nanostructures with uniform properties. In this context, the bottom-up approach, *i.e.* elaborating nanostructures in a self-ordered manner and integrating them directly into devices, seems to be a promising way to push towards miniaturization of microelectronics components and to create new functionalities.[1]

In this chapter, we will focus our attention on two emerging nanotechnologies we have developed to fabricate, organize and integrate nanomaterials (nanodots, nanowires) in nanoelectronics devices.

The first one uses the natural properties of diblock copolymers[2] to create (i) a hexagonal lattice array of vertical PMMA cylinders in a polystyrene (PS) matrix, or (ii) an ordered array of horizontal PMMA cylinders in a PS matrix. These films are then used as deposition and/or etching masks to produce dense lattices of 20 nm diameter nanoparticles and nanowires with a density of around 10^{11} cm^{-2}. Examples of integration with an 8-inch wafer memory technology will be shown.

The second technology is based on chemical vapor deposition (CVD)[3] of basic nanomaterial building blocks for future microelectronics. We focus on CVD growth of Si nanostructures using a fully compatible CMOS process. The resulting nanoobjects are integrated in test devices and measured electrically.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

2. Diblock copolymer technology for nano-objects fabrication

In this chapter, we use an asymmetric polystyrene-block-methyl methacrylate (PS-b-PMMA) diblock copolymer having a PMMA weight fraction of 0.3 with a molecular weight (MW) of 67100 g·mol^{-1} and a polydispersity of 1.09. A thin ~39 nm film of PS-b-PMMA was deposited on the substrate by spin-coating a toluene-diluted solution, followed by annealing at 170 °C for 24 hours. In order to employ this film as a template for nanodot or nanowire fabrication, the samples were exposed to ultraviolet (UV) light (HgXe with a ~1.2 J/cm^2 fluence) for 15 minutes. The PMMA is degraded via a chain scission under UV irradiation whereas the PS is cross-linked and becomes insoluble. Rinsing with acetic acid then removes the PMMA phase, leaving an array of holes in the PS matrix. Then, a 5 nm thick layer of metal, Pt for example, was deposited onto the deposition mask by dc magnetron sputtering under an Ar pressure of 2.0 mT at ambient temperature. The formation of self-organized nanodots and nanowires is obtained by a lift-off process based mainly on O_2 plasma. The nanodots could for example be integrated in a capacitor or served as a hard mask to fabricate an ordered array of nanopillars.[4]

This technology was applied to form nanopillars in a multilayer made of a $Si/Si_{0.8}Ge_{0.2}$ 10/10 nm superlattice epitaxially grown on a Si (100) substrate by reduced pressure CVD. Here, the Pt nanodots provided a hard mask and Cl_2/O_2 inductively coupled plasma (ICP) was used to anisotropically etch our Si/SiGe stacks. The entire processing sequence is schematically illustrated in Fig. 1.

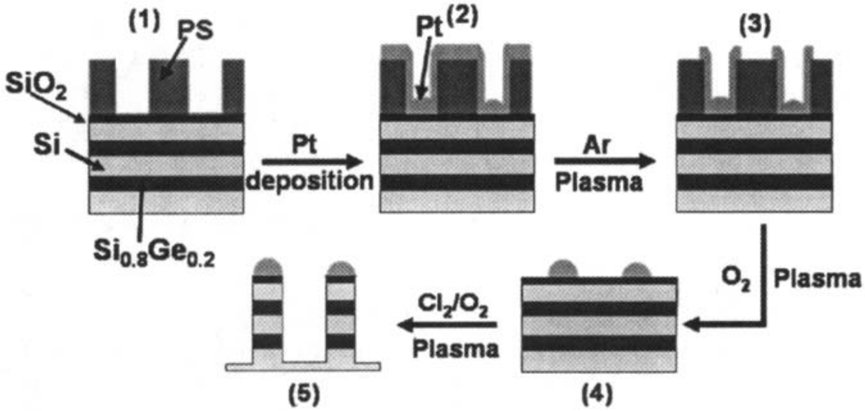


Figure 1. Schematic process flow for Pt dot fabrication by lift-off technique and realization of pillars by using Pt dots as a hard mask: (1) PS porous template on a $SiO_2/Si/Si_{0.8}Ge_{0.2}/Si/Si_{0.8}Ge_{0.2}/Si$ multilayer; (2) Pt deposition onto the mask; (3) Ar plasma to selectively remove the metal in excess on top of the PS layer; (4) O_2 plasma to remove the residual polymer and side-walls metal-coated PS; and (5) etching the multilayer with a Cl_2/O_2 ICP plasma.

Figure 2 shows the nanopillars etched from a two-layer Si/$Si_{0.8}Ge_{0.2}$ stack with a Cl_2/O_2 (180 sccm/25 sccm) ICP plasma during 15 s. The nanopillars have a cylindrical shape with the same height (~50 nm) and diameter (~20 nm). The Pt dots are still present after the etching step. This means that the increase of the O_2 flow rate in the etching gas limits the lateral etching of $Si_{0.8}Ge_{0.2}$ quantum dots. Nevertheless, Fig. 2 shows that the diameter of the pillar is lower at the locations of the $Si_{0.8}Ge_{0.2}$ layers. The significant decrease of the lateral etching of $Si_{0.8}Ge_{0.2}$ layers with the increasing O_2 flow is due to sidewall passivation by an oxide layer.[5] The thinner diameter of the $Si_{0.8}Ge_{0.2}$ regions of the nanopillars is attributed to a thinner passivation layer than on the Si regions of the nanopillars. This effect was also observed by Monget *et al.*[5] By analyzing the composition of the passivation layer by x-ray photoelectron spectroscopy (XPS), they determined that it contained Si, O, and Cl. As no Ge could be detected, they concluded that germanium etch products desorb into the gas phase without re-depositing on the sidewalls. This means that only silicon contributes to formation of the passivation layer, which is consequently thinner on the SiGe layers than on the Si layers.

In order to access to the physico-chemical nature of the nanopillars after the etching process, XPS measurements were performed on the samples (not shown). The Si spectrum shows two contributions at 99.6 eV and 104 eV, commonly attributed to Si-Si and Si-O bonds, respectively. The Ge 2p electronic level shows Ge-Ge and Ge-O contributions at 1218 eV and 1221.5 eV respectively.[6] An interesting point to note here is that the 1221.5 eV peak corresponds to a completely oxidized Ge state[7] with no significant sub-oxide. The presence of the Ge-Ge and Si-Si peaks shows that the Ge structure of the nanopillar is still present

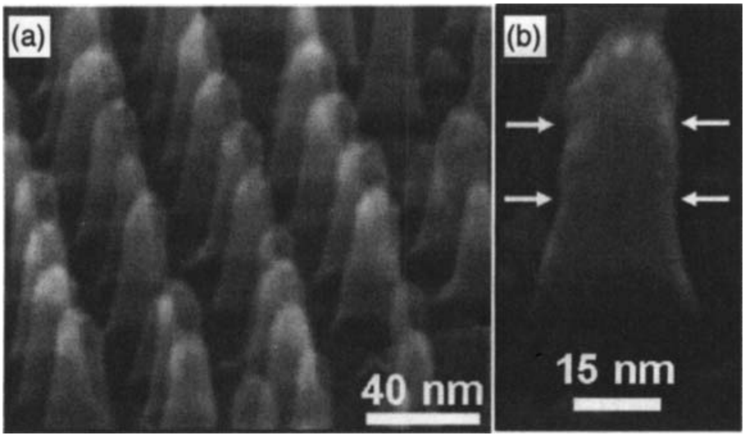


Figure 2. SEM images of pillars obtained by using the Pt dot array as an etch mask in a Cl_2/O_2 ICP etch process of Si/$Si_{0.8}Ge_{0.2}$ 10/10 nm two-period superlattice on a Si substrate: (a) multiple pillars; (b) isolated ~50 nm tall pillar. The arrows in (b) indicate that the diameter of the SiGe layers is lower due to lateral etching.

after the etching, whereas the Ge-O and Si-O contributions are likely due to the sidewall passivation layers. The high reactivity of the oxygen radicals present in the gas phase during ICP explains the completely oxidized Ge state observed here.

3. Chemical vapor deposition of nanodots and nanowires

During the last few years, silicon quantum dots (Si-QDs) have been studied for nanoelectronics applications. Their unique physical properties, due to size confinement effects and Coulomb blockade phenomena, make Si-QDs suitable for use in new silicon based devices like single electron transistors[8] or quantum dot floating gate memories.[9] Low pressure CVD is a good way to obtain Si-QDs for industrial applications because of its MOSFET technology compatibility. By controlling the early stages of the Si film growth, silicon crystallites of nanometer size (5 nm) are obtained – a size that has proven of interest to nonvolatile memory devices.[10] In order to obtain reliable devices, size, size uniformity and Si-QDs density must be controlled with a great precision and reproducibility. We have shown that a two-step CVD process makes it possible to control the QD size distribution while obtaining densities up to 10^{12} cm^{-2} – see Fig. 3.

Traditionally, Si nanowires (NWs) are grown using the vapor-liquid-solid (VLS) technique[11] where a metallic catalyst aids in the decomposition of a Si-containing gas. The metal and Si form a liquid eutectic droplet that moves along with the growing end of the wire as solid Si precipitates from the melt. Gold is the most frequently used catalyst because the Au-Si eutectic point occurs at a low temperature (363 °C).[12] We show that Si NWs can be grown with a solid-phase epitaxy process using fully CMOS-compatible catalysts: PtSi, NiSi, and PdSi.

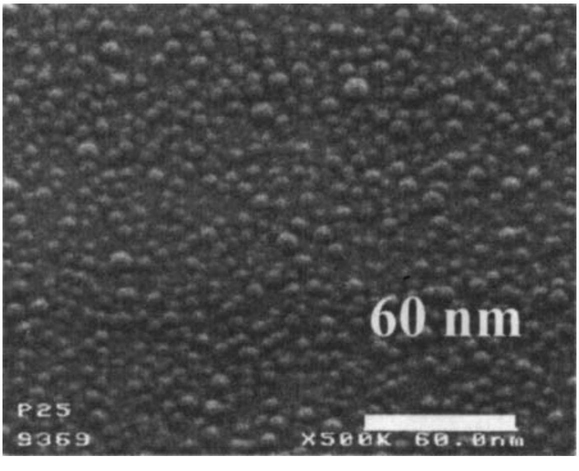


Figure 3. Typical SEM picture of Si nanocrystals after deposition on a SiO_2 surface: QD density is ~10^{12} cm^{-2}, mean QD diameter is ~8 nm.

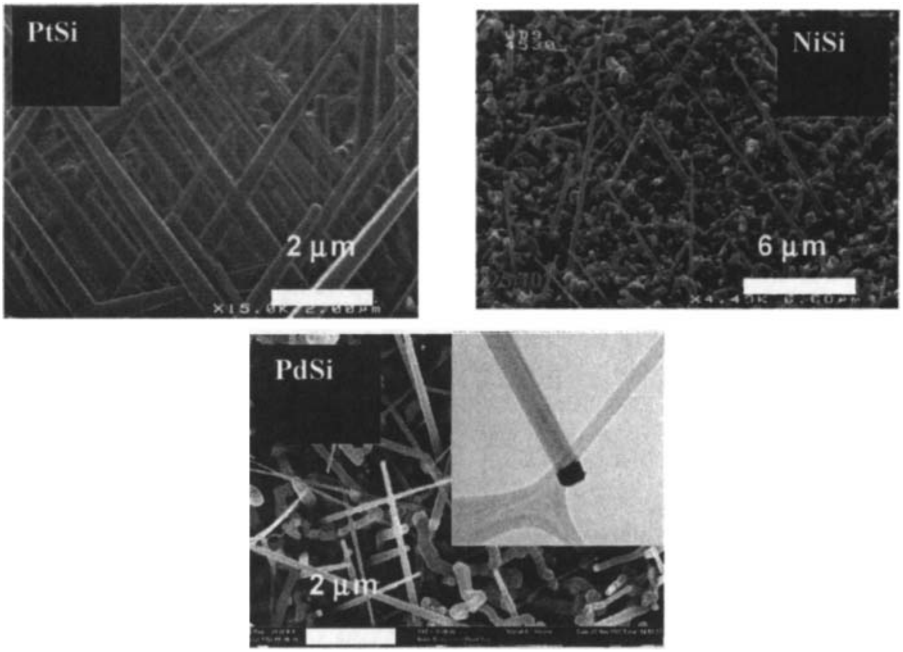


Figure 4. SEM picture of Si NWs obtained with different silicide catalysts: PtSi, NiSi and PdSi. The inset in the PdSi-catalyzed NW SEM shows a TEM analysis confirming the crystalline nature of the NWs.

Thus we have grown silicon nanowires by CVD using silicide catalysts and SiH_4 diluted in H_2 as the precursor gas. The resulting NWs are shown in Fig. 4. The growth temperature varied between 400 and 1000 °C, with silane fluxes in the 10–50 sccm range. Extra HCl gas could be added to favor catalytic growth.

4. Integration of nanoobjects

We have deposited Pt nanoparticles on *p*-type Si (100) substrate covered by a thin tunnel silicon dioxide (5 nm) via the diblock copolymer technology. This layer is integrated as a floating gate in a CMOS capacitor on an 8-inch pilot CMOS fab line. The density of nanoparticles was around 10^{11} cm^{-2} with the mean size centered around 20 nm. The Pt particles were covered by a 20 nm layer of SiO_2 deposited at low temperature. Aluminum was used as a metal gate. Figure 5 presents pictures representing the different steps of the process.

Capacitance-voltage characterization has been performed for devices with and without Pt nanodots. After a gate stress of 12 V during for 500 ms, a flatband voltage shift of about 1 V was only observed in devices with nanoparticles, indicating charge storage in the Pt nanocrystals.

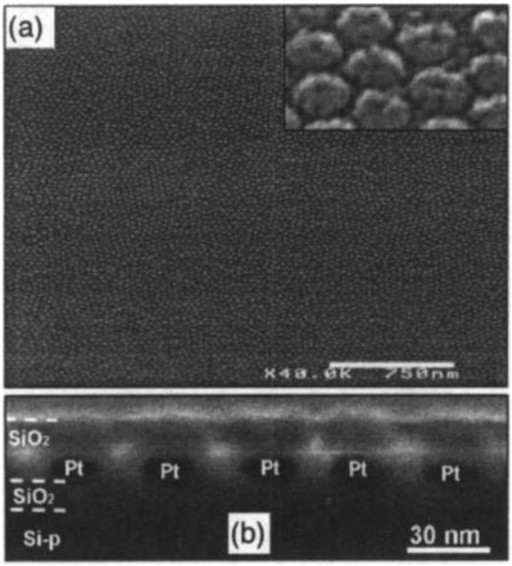


Figure 5. SEM pictures of the organized Pt nanocrystals: (a) top view; (b) cross-sectional view after the low temperature SiO_2 deposition.

Silicon NW field-effect transistors (FETs) were fabricated on thermal SiO_2 which is used as gate oxide and heavy *p*-doped Si substrate as back gate electrode – see Fig. 6. The source/drain electrodes were defined using photolithography, Ni/Al 120 nm/50nm metal deposition and lift-off. Through a controlled nickel silicide, the embedded parts of the undoped NWs are tuned into nickel silicide NWs, thus providing a direct metallic nanolead to the semiconducting nanowires and enhancing the device performance. The transistors obtained with NWs of ~100 nm diameter display *p*-type behavior, current densities around 2 kA/cm^2 and I_{ON}/I_{OFF} ratio up to 10^7.

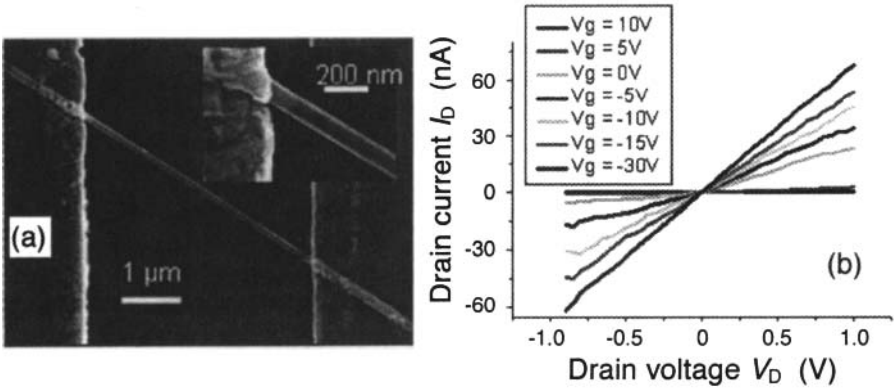


Figure 6. (a) SEM image of Si NW FET, the inset shows the axial silicidation of the Si NW; (b) output characteristics of the back-gated Si NW FET.

5. Conclusions

We believe that the fabrication technologies based on the bottom-up approach are promising to push forward the limits of conventional technology. Ordered arrays of dense nanostructures could be obtained and integrated in devices with CMOS process compatibility. Adding new functionalities to a CMOS integrated circuit via a 3D integration technology could enable new bottom-up applications.

Acknowledgments

This work was supported by the ANR French agency project PREAANS.

References

1. T. Baron, F. Martin, P. Mur, *et al.*, "Low pressure chemical vapor deposition growth of silicon quantum dots on insulator for nanoelectronics devices," *Appl. Surf. Sci.* **164**, 29 (2000).
2. P. Mansky, Y. Liu, Y. L. Huang, T. P. Russell, and C. Hawker, " Controlling polymer-surface interactions with random copolymer brushes," *Science* **275**, 1458 (1997).
3. T. Baron, F. Mazen, J. M. Hartmann, *et al.*, "Growth and characterization of LPCVD Si quantum dots on insulators," *Solid State Electronics* **48**, 1503 (2004).
4. K. Aissou, T. Baron, M. Kogelschatz, *et al.*, "Fabrication of well-organized and densely packed Si nanopillars containing SiGe nanodots by using block Copolymer Templates," *Chem. Mater.* **20**, 6183 (2008).
5. C. Monget, S. Vallon, F. H. Bell, L. Vallier, and O. Joubert, "X-ray photoelectron spectroscopy analyses of oxide-masked polycrystalline SiGe features etched in a high-density plasma source," *J. Electrochem. Soc.* **144**, 2455 (1997).
6. J. F. Moulder, W. F. Stickle, P. E. Sobol, and K. D. Bomben, *Handbook of X-Ray Photoelectron Spectroscopy*, 2nd ed., Eden Prairie, MN: Perkin-Elmer Corporation, 1992.
7. B. Pelissier, H. Kambara, E. Godot, E. Veran, V. Loup, and O. Joubert, "XPS analysis with an ultra clean vacuum substrate carrier for oxidation and airborne molecular contamination prevention," *Microelectronic Eng.* **85**, 151 (2008).
8. B. H. Choi, S. W. Hwang, I. G. Kim, H. C. Shin, Y. Kim, and K. Kim, "Fabrication and room-temperature characterization of a silicon self-assembled quantum-dot transistor," *Appl. Phys. Lett.* **73**, 3129 (1998).

9. M. Kanoun, C. Busseret, A. Poncet, A. Souifi, T. Baron, and E. Gautier, "Electronic properties of Ge nanocrystals for nonvolatile memory applications," *Solid State Electronics* **50**, 1310 (2006).
10. B. De Salvo, G. Ghibaudo, P. Luthereau, *et al.*, "Transport mechanisms and charge trapping in thin dielectric/Si nano-crystals structures," *Solid State Electronics* **45**, 1513 (2001).
11. R. S. Wagner and W. C. Ellis, " Vapor-liquid-solid mechanism of single crystal growth," *Appl. Phys. Lett.* **4**, 89 (1964).
12. M. S. Gudiksen, L. J. Lauhon, J. Wang, D. C. Smith, and C. M. Lieber, "Growth of nanowire superlattice structures for nanoscale photonics and electronics," *Nature*, **415**, 617 (2002).

Scrolled Si/SiGe Heterostructures as Building Blocks for Tube-Like Field-Effect Transistors

N. Demarina[*] **and D. Grützmacher**
IBN-1, Forschungszentrum Jülich, D-52425 Jülich, Germany

1. Introduction

Microelectronic technology has entered almost every segment of our lives. Today, there are microprocessors in more than 20% of consumer products and it is expected that the share will increase with ongoing miniaturization. However, with increasing demand for higher performance, it is expected that mainstream technology will branch off aiming for speed, functionality, computing capacity or packaging density, depending on the application. Currently and certainly for another 10 years to come, the rapid improvement in computing performance will be achieved by the continued downscaling of dimensions of Si-based components. The downscaling of complimentary-metal-oxide-semiconductor (CMOS) devices decreases their capacitance, resulting in higher circuit operating speeds and lower power consumption. Further, downscaling increases the number of components in the circuit and enhances parallel operation capability, resulting in another increase in circuit performance. Recent CMOS downscaling trends have been very aggressive. Advanced CMOS technology has already reached gate length L_G of 65 nm, Intel recently announced volume production of 45 nm technology, and working transistors with L_G = 6 nm have been reported.[1] This gate length is only twenty times of the atomic distance between Si atoms. Thus, it is certainly true that the downscaling of the components is approaching its limit. At the same time the power density on the chip has been increasing dramatically, making heat management into another challenging problem for CMOS circuitry.

On the evolutionary way towards nanoelectronics, new functional devices made from SiGe nanoobjects promise an exciting but challenging era. Large-scale manufacturing of these devices is predicated on several key requirements: controllability of all parameters, high-precision positioning, reliability, high performance, ultralow switching power and low cost. A number of SiGe-based nanodevices have been proposed, including like ultrathin-body GeOI transitors,[2] the Ge dot-FETs,[3] and spintronic devices using *g*-factor control via the Ge concentration in Si/SiGe quantum structures[4]. Thus, the fabrication and physics of SiGe nanoscale objects are of great interest to future nanoelectronics.

---

[*] Also with Nizhniy Novgorod State University, 603950 Nizhniy Novgorod, Russia.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

The downscaling of nanoelectronics devices puts severe constrains on the uniformity of devices, since on the nanoscale statistical fluctuations in geometrical dimensions or doping concentrations might be too large for proper circuit design. In addition, the classical top-down approach to the fabrication of nanostructures via lithography and etching leads to surface damage, which might be detrimental for device operation. The latter problem can be elegantly circumvented using self-organization for fabrication of nanostructures.

The invention of the self-scrolling technique by Victor Prinz and co-workers[5] has been immediately employed for fabrication of a novel class of nanoobjects. Several groups in the world picked up the technology, applying it to various strained material systems, such as InGaAs/GaAs[6-10], Si/SiGe[11-13], and InGaP[14] to fabricate the freestanding micro- and nanotubes. So far, not only bilayers, but also multilayer systems[15] as well as a single-material layers[12] have been transformed into scrolled objects. The formation of tubes, rings, and coils has been achieved also by releasing Si/SiGe, SiGe/ Si, and Si/SiGe/Si[13,16] opening ways to integrate novel functional devices into the CMOS technology. Fabrication of rolled-up shells with arrays of embedded quantum dots is also of considerable interest due to modification of electronic and optical properties of the shells.[17]

Micro- and nanotubes are also excellent objects for studying properties of two-dimensional electron gas (2DEG) on cylindrical surfaces, including charge-carrier ballistic transport and spin-orbital interaction. A theoretical study of quantum properties of InAs/GaAs nanotubes revealed spatial separation of the charge carriers (electrons and holes) in nanotube walls.[18] First experiments on 2DEG magnetotransport properties in rolled-up quantum wells were reported in Refs. 19 and 20. Subsequently, more studies were published on similar GaAs-based structures with different contact configurations.[21-23] Magnetotransport properties of the rolled-up 2DEG are affected by the inhomogeneous normal component of the magnetic field, as well as by the curved surface itself. The remarkable effect of the magnetic field gradient arising from the cylindrical surface curvature on magnetotransport of high-mobility 2DEG was reported in Ref. 24. To date, neither two-dimensional hole gas (2DHG) in rolled-up shells nor 2D systems of charge carriers in Si/SiGe rolled-up shells have been reported. Besides the theory of magnetotransport published recently[25] we are familiar with only one more theoretical study[26] that revealed strong anisotropy of spin relaxation times. Properties of high-mobility hole and electron gas in the rolled-up structures subject to only an electric field have not been considered so far. Furthermore, Si/SiGe scrolled structures are of particular interest for transport studies, since it is expected that using well-defined scrolling directions, the applied shear strain will dramatically reduce the effective mass for electrons in the strained Si.[27]

In this chapter, we focus on the Si/SiGe rolled-up heterostructures grown on a Si substrate. In this type of structures, a large band offset is mainly restricted to the valence band, making them useful for *p*-type devices. We analyze the dependence of hole density in the scrolled structures on the doping level and surface state density. We also present a study of the low-field hole mobility in the rolled structures in an electric field parallel to the structure axis.

2. Charge carrier density in modulation doped scrolled $Si/Si_{1-x}Ge_x$ structures

We consider a scrolled structure[28] formed from a bilayer consisting of Si and SiGe layers of h_{Si} and h_{SiGe} thickness, correspondingly, grown on (001) Si. In our theoretical treatment we assume that the bilayer structure is rolled into a single-wall cylinder. The doping is applied to the layer of Si located at the inner surface of the tube with the width of 2 nm as it is shown in the inset of Fig. 1. This model of the doping approximately corresponds to delta-doping of the original structure.

For the calculation of the hole density and mobility in the structure we first determine the strain in a stationary configuration of the tube making use of the linear elasticity theory. Following Grundmann,[29] strain components in the tube grown on (001) oriented substrate along r, φ, and z directions, denoted as ε_r, ε_φ and ε_z respectively, can be written as functions of the tube internal radius R and the lattice constant on the inner Si surface of the tube a_{in} as follows:

$$\varepsilon_\varphi^{(j)}(r) = a_{in}(r+R)/a_0^{(j)} - 1;\ \varepsilon_z^{(j)}(r) = a_{Si}/a_0^{(j)} - 1;\ \varepsilon_r^{(j)}(r) = -C_{12}^{(j)}(\varepsilon_\varphi^{(j)}+\varepsilon_z^{(j)})/C_{11}^{(j)},$$

where j labels the material (Si or Ge_xSi_{1-x}); $a_0^{(j)}$ and a_{Si} denote unstrained lattice constants of the material (j) and the Si substrate; and $C_{11}^{(j)}$ and $C_{12}^{(j)}$ are the elastic coefficients.

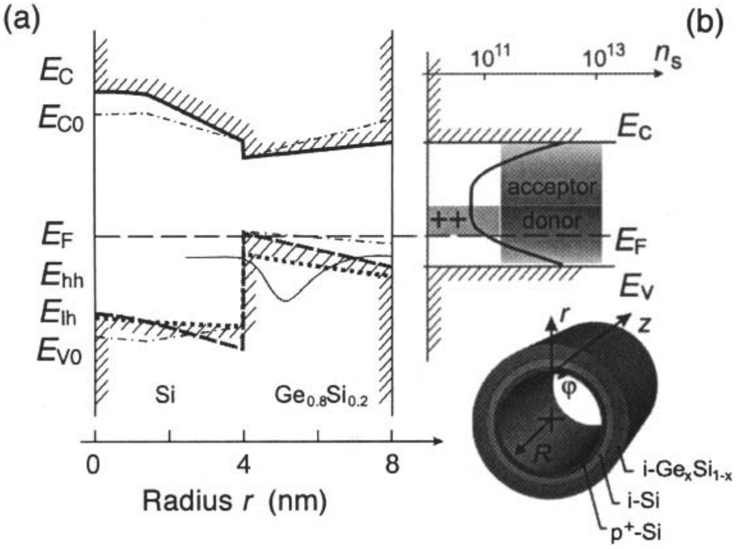


Figure 1. (a) Conduction (E_C) and valence (E_V) bands and Fermi level E_F in the $Si/Si_{0.2}Ge_{0.8}$ scrolled structure calculated neglecting (thin dot-dashed line) and taking into account (thick lines) the surface states and doping; dashed and dotted lines indicate position of the Γ-point of the heavy- and light-hole states correspondingly; the width of the both Si and $Si_{0.2}Ge_{0.8}$ layers is 4 nm; (b) surface state density n_S at the SiO_2/p-Si interface. Inset: sketch of a scrolled Si/SiGe heterostructure consisting of Si and SiGe layers.

Minimization of the total elastic energy, calculated by integrating the elastic energy density $w = (C_{11} - C_{12})[(C_{11} + C_{12})(\varepsilon_\varphi^2 + \varepsilon_z^2) + 2C_{12}\varepsilon_z\varepsilon_\varphi]/2C_{11}$ over the tube volume, makes it possible determine the values of the radius and lattice constant of the tube in the equilibrium state. We note that the radius value calculated using this model is in a good agreement with the available experimental data.[30]

The strain distribution strongly affects position of the conduction and valence band. Hydrostatic strain shifts the energy position of bands, while biaxial strain splits the degenerate bands.[31] First, we calculate the position of the conduction (E_{C0}) and valence (E_{V0}) band edges in scrolled structures taking into account only the strain contribution to the band bending following the expressions from Ref. 31. The dash-dotted line in Fig. 1(b) shows the calculated position of the valence and conduction band edges for the $Si/Si_{0.2}Ge_{0.8}$ scrolled structure. The band offset is mainly restricted to the valence band and the $Si_{0.2}Ge_{0.8}$ layer is a potential well for holes. Second, we take into account the additional electrostatic potential $V(r)$ due to the electric charge arising from modulation doping, as well as the charged surface states formed after the oxidation of the Si and $Si_{1-x}Ge_x$ surfaces. Therefore, the position of the conduction and valence bands fulfils $E_{C(V)} = E_{C(V)0} - qV(r)$, where q denotes the positive elementary charge. We find the electrostatic potential by the self-consistent solution of Poisson and Schrödinger equations, where the Schrödinger equation is solved for the envelope functions within the effective mass approach for both heavy- and light-hole states. We note that for the typical layer thicknesses of the scrolled structures, the radius exceeds 50 nm, which eliminates quantization of the hole spectrum for the momentum quantum number. As a result, the holes in rolled structures effectively form a 2DHG. We integrate the Poisson equation neglecting the free electron charge and assuming that the doping atoms are fully ionized and the hole density is determined by the corresponding wave functions and occupation of the energy states. The total density of charged surface states at the inner surface determines the magnitude of the electric field normal to the surface, providing the required boundary condition for the Poisson equation. The full band bending is determined by the condition that the total space charge in a scrolled structure is equal and opposite to the charge accumulated by the surface states.

We describe the density of surface states on the inner and external structure surfaces as a function of energy W by the well-known U-type dependence, which for the (001) Si surface after oxidation at room temperature is presented in Ref. 32. We fit the experimental data with the following expression:

$$N_S(W) = A_{NS}\cosh[B_{NS}(W - W_{mid})^2]\,,$$

where W_{mid} denotes the midgap energy in eV, $A_{NS} = 5\times10^{11}\ \mathrm{cm^{-2}eV^{-1}}$, and $B_{NS} = 18\ \mathrm{eV^{-2}}$. The charge of the surface states depends on the type of the state and its occupation by electrons. Note that, in accordance with the conventional model,[33] the states lying above the mid gap are acceptor-type and below are donor-type. Figure 1(a) displays the calculated position of the bands for the $Si/Si_{0.2}Ge_{0.8}$ structure with the p-doping density in the Si layer of $10^{12}\ \mathrm{cm^{-2}}$ and the total surface density at the each surface of $10^{12}\ \mathrm{cm^{-2}}$. The corresponding surface state density at the Si surface is presented in Fig. 1(b). In this particular case, the Fermi level near the surfaces is approaching the valence band edge, leading to positive charge on both Si and SiGe interfaces. This, together with the modulation doping, gives rise to the downward tilt of the bands near the $Si_{0.2}Ge_{0.8}$ layer surface and the resulting accumulation of the hole gas at the $Si/Si_{0.2}Ge_{0.8}$ interface.

To reveal the influence of the surface states on the density of free holes in the channel, we vary the total surface state density by changing the coefficient A_{NS} within the 10^8–10^{12} $cm^{-2}eV^{-1}$ range, without altering the qualitative shape of the surface state density function. We consider the Si and SiGe surfaces separately, first, increasing the density of surface states at the SiGe surface and assuming the Si surface to be ideal in the sense that the total surface state density is about 10^{10} cm^{-2}, and then *vice versa*.

Figure 2(a) shows the hole density in the $Si/Si_{1-x}Ge_x$ structure with the width of both layers set to 4 nm and *p*-doping density of the Si layer set to $5x10^{18}$ cm^{-3}. The hole density decreases with increasing surface state density, as more holes are trapped on the surface states. The surface states at the SiGe surface affect the hole density more strongly than those located at the Si surface. For a surface state density at the SiGe surface above $2x10^{12}$ cm^{-2}, the SiGe channel is completely depleted at low temperature and the scrolled structure becomes nonconductive,

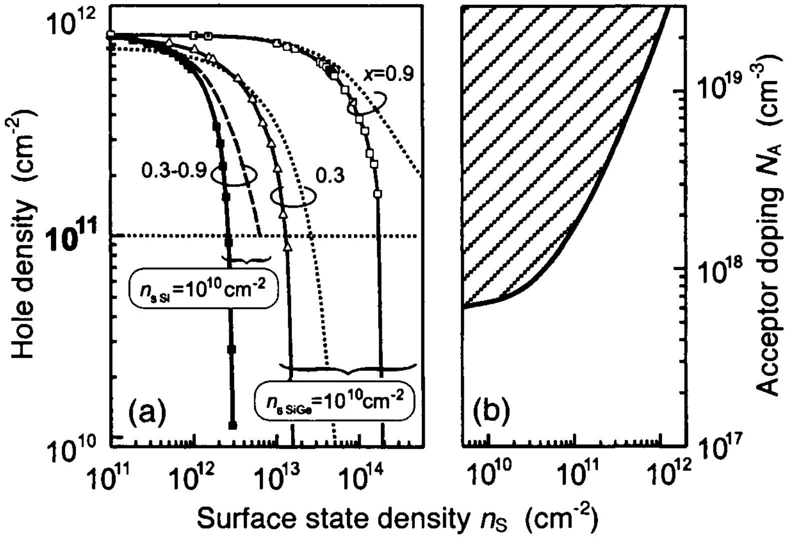


Figure 2. (a) Hole density in the $Si_{1-x}Ge_x$ layer *vs.* density of surface states n_S at $Si_{1-x}Ge_x$ and Si surfaces for different Ge content values (x) in the $Si_{1-x}Ge_x$ layer calculated for low (solid line) and room (dashed and dotted lines) temperature; (b) shaded area shows required doping acceptor density N_A in the p^+-Si layer *vs.* density of surface states n_S at the Si and $Si_{1-x}Ge_x$ surfaces needed to obtain a hole density in the SiGe layer above 10^{11} cm^{-2}.

while at the Si surface the surface state density might reach above 10^{13} cm^{-2} with the hole density remaining above 10^{11} cm^{-2}. This we attribute to the relatively close position of E_F to the valence band edge at the surface of the SiGe layer, where the surface state density function rises more steeply than in the middle of the gap – see Fig. 1(b). Due to the additional hole activation at room temperature, the hole channel remains conductive for the higher value of surface state density at both surfaces.

So far, we have discussed the hole density as a function of the surface state density while keeping the doping fixed. We have also calculated the minimal doping level required to produce a hole density above 10^{11} cm^{-2} in the $Si_{1-x}Ge_x$ channel, making the tubes conductive. Thus, we consider $Si/Si_{1-x}Ge_x$ scrolled structures with x varying from 0.3 to 0.9 and the width of the both layers from 2 to 12 nm. The layer thicknesses are restricted by the critical thickness for pseudomorphic strained layers.[31] We increase the total surface state density from 5×10^8 to 10^{12} cm^{-2}, where the lower value corresponds to the surface that is additionally passivated with hydrogen, while the higher value corresponds to a surface exposed to the oxidation in air at room temperature. We also assume that the density of states on SiGe surface is larger than on the Si surface by a factor of 5. The shaded area in Fig 2(b) shows the doping density that is required for a given surface state density to keep the hole density in the channel of the scrolled structure above 10^{11} cm^{-2}. The calculation shows that the chosen 2DHG density can be reached for any configuration of the tube within the defined limits by means of available *p*-doping. We note that additional passivation of the surface states makes it possible to reach higher conductivity of the SiGe channel.

3. Low-field hole mobility in rolled-up Si/SiGe structures

We calculate the low-field hole mobility in the scrolled structures subject to a static electric field applied along the structure axis at low temperature by means of the three-dimensional single-particle Monte Carlo method.[34] We treat inelastic hole scattering by optical phonons and elastic scattering by acoustic phonons via deformation-potential coupling for bulk material. We also take into account interface roughness by using a scattering potential given by a δ-function at a perfect interface,[35] assuming small thickness fluctuations of 0.4 nm and a surface roughness correlation length in the direction parallel to the surface of about 2.5 nm.[36] The scattering rates for every structure are calculated numerically using the hole wave functions and energy levels obtained from the coupled Poisson and Schrödinger equations.

In the hole transport simulations, we always assume that the surface state density at the both structure surfaces is about 10^{10} cm^{-2} and choose the modulation doping level to provide a hole density in the SiGe channel of about 10^{11} cm^{-2}. Figure 3(a) shows the hole mobility as a function of Ge content in the $Si/Si_{1-x}Ge_x$ rolled structure with both layers set to a range of thicknesses: 2, 4, 6 and 10 nm.

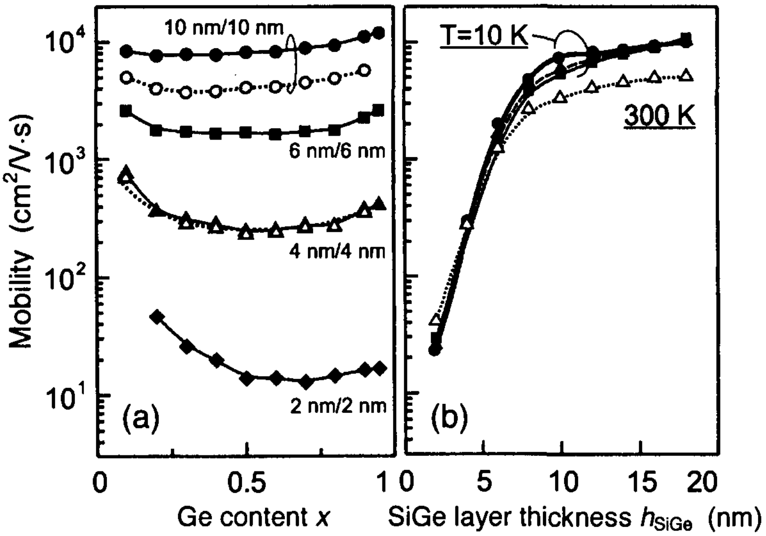


Figure 3. (a) Low-field hole mobility along the rolled structure axis *vs.* Ge content x in the $Si_{1-x}Ge_x$ layer; the thicknesses of the Si and $Si_{1-x}Ge_x$ layers are equal and set to 2, 4, 6, and 10 nm; (b) low-field hole mobility along the $Si/Si_{0.5}Ge_{0.5}$ structure axis *vs.* the $Si_{0.5}Ge_{0.5}$ layer thickness. The calculations are performed for low (solid lines) and room (dotted lines) temperature.

The mobility dependence displays a pronounced minimum for the value of the Ge content $x \sim 0.5$. We attribute the initial decrease in mobility with x at low $x < 0.5$ to hole scattering by the Si/SiGe interface roughness, which dominates other scattering processes at low temperature and becomes stronger as the Ge increases. Conversely, the substantial reduction of the hole effective mass in the SiGe layer with the larger Ge content ensures increasing mobility for $x > 0.5$. We note that the mobility also substantially drops with a decrease in layer thickness, which can be attributed to the stronger penetration of the hole wave function tail into the Si layer, leading to the stronger interplay between holes and surface imperfections. The presence of strong interface scattering in rolled structures with thin layers is also validated by the mobility calculation performed at room temperature, shown in Fig. 3(a). In scrolled structures with thick SiGe wells (above 8 nm), at room temperature the holes are also effectively scattered by optical phonons, yielding a mobility lower than the low temperature value by a factor of ~2. In structures with narrow quantum wells, interface roughness scattering always dominates and makes the mobility temperature-independent.

Figure 3(b) shows that the low-field hole mobility increases with the width of the $Si_{0.5}Ge_{0.5}$ layer when the Si layer is set to 4, 6, and 10 nm. As we have already discussed above, interface roughness scattering becomes less effective as the SiGe layer thickness increases, leading to a substantial increase in mobility from 20 to 1000 cm^2/V·s. The width of the Si layer affects the mobility value only slightly, as the holes are manly confined to the SiGe layer. Comparison of the mobility calculated for low and room temperature enables us to conclude that the optical phonon scattering affects hole transport in the scrolled structures only if the SiGe layer thickness is larger than 5 nm. In this respect, it is important to note that the strain relaxation during the scrolling process may enable a technology to substantially overcome the limits imposed by the critical thickness of SiGe on Si in planar technology.

4. Conclusions

This chapter reports a Monte Carlo study of the low-field hole transport in scrolled Si/SiGe heterostructures. We find that interface roughness scattering substantially reduces the hole mobility and is the main scattering mechanism in the SiGe scrolled structures. However, an increase in the width of the SiGe layer is accompanied by a marked increase in the hole mobility, which reaches the value of 10^4 cm^2/V·s. The study delivers criteria for development of a building block for a tube-like field-effect transistor.

Acknowledgments

This work was partially supported by the DAAD, by the Russian Ministry of Science and Education via the Mikhail Lomonosov program and Grant No. BF5M01 of the Russian Ministry of Science and Education, and by the U.S. Civilian Research and Development Foundation.

References

1. B. Doris, M. Ieong, T. Kanarsky, *et al.*, "Extreme scaling with ultra-thin SOI channel MOSFETs," *Tech. Digest IEDM* (2002), p. 267.
2. A. Zaslavsky, S. Soliveres, C. Le Royer, S. Cristoloveanu, L. Clavelier, and S. Deleonibus, "Negative transconductance in double-gate germanium-on-insulator field effect transistors," *Appl. Phys. Let.* **91**, 183511 (2007).
3. O. G. Schmidt, A. Rastelli, G. S. Kar, *et al.*, "Novel nanostructure architectures," *Physica E* **25**, 280 (2004).
4. H. Malissa, D. Gruber, D. Pachinger, F. Schäffler, W. Jantsch, and Z. Wilamowski, "Design and fabrication of a SiGe double quantum well structure for g-factor tuning," *Superlatt. Microstruct.* **39**, 414 (2006).
5. V. Ya. Prinz, V. A. Seleznev, A. K. Gutakovsky, *et al.*, "Free-standing and overgrown InGaAs/GaAs nanotubes, nanohelices and their arrays," *Physica E* **6**, 828 (2000).
6. O. Schmidt, N. Schmarje, Ch. Deneke, C. Müller, and N.-Y. Jin-Phillipp, "Three-dimensional nano-objects evolving from a two-dimensional layer technology," *Adv. Mater.* **13**, 756 (2001).

7. O. G. Schmidt and K. Eberl, "Thin solid films roll up into nanotubes," *Nature* **410**, 168 (2001).
8. O. Schmidt, C. Deneke, N. Schmarje, C. Müller, and N. Y. Jin-Phillipp, "Free-standing semiconductor micro- and nano-objects," *Mat. Sci. Eng. C* **19**, 393 (2002).
9. C. Deneke, C. Mueller, N. Jin-Phillipp, and O. G. Schmidt, "Diameter scalability of rolled-up In(Ga)As/GaAs nanotubes," *Semicond. Sci. Technol.* **17**, 1278 (2002).
10. D. J. Bell, L. Dong, B. J. Nelson, M. Golling, L. Zhang, and D. Grützmacher, "Fabrication and characterization of three-dimensional InGaAs/GaAs nanosprings," *Nano Lett.* **6**, 725 (2006).
11. R. Songmuang, N. Y. Jin-Phillipp, S. Mendach, and O. G. Schmidt, "Single rolled-up SiGe/Si microtubes: Structure and thermal stability," *Appl. Phys. Lett.* **88**, 21913 (2006).
12. R. Songmuang, Ch. Deneke, and O. G. Schmidt, "Rolled-up micro- and nanotubes from single-material thin films," *Appl. Phys. Lett.* **89**, 223109 (2006).
13. F. Cavallo, R. Songmuang, C. Ulrich, and O. G. Schmidt, "Rolling up SiGe on insulator," *Appl. Phys. Lett.* **90**, 193120 (2007).
14. O. Schmidt, C. Deneke, Y. Nakamura, *et al.*, "Nanotechnology – Bottom-up meets top-down," *Adv. Sol. State Phys.* **42**, 231 (2002).
15. C. Deneke, U. Zschieschang, H. Klauk, and O. G. Schmidt, "InGaAs/GaAs/alkanethiolate radial superlattices," *Appl. Phys. Lett.* **89**, 263110 (2006).
16. M. Huang, C. Boone, M. Roberts, *et al.*, "Nanomechanical architecture of strained bilayer thin films: from design principles to experimental fabrication," *Adv. Mater.* **17**, 2860 (2005).
17. T. Kipp, H. Welsch, Ch. Strelow, Ch. Heyn, and D. Heitmann, "Optical modes in semiconductor microtube ring resonators," *Phys. Rev. Lett.* **96**, 0774039 (2006).
18. V. M. Osadchii and V. Ya. Prinz, "Charge-carrier separation in rolled heterostructures," *JETP Lett.* **72**, 312 (2000).
19. A. B. Vorob'ev, V. Ya. Prinz, and A. I. Toropov, *Proc. Intern. Conf. Superlatt. Nanostruct. Nanodev.*, Toulouse, France, 2002.
20. A. B. Vorob'ev, V. Ya. Prinz, Yu. S. Yukecheva, and A. I. Toropov, "Magnetotransport properties of two-dimensional electron gas on cylindrical surface," *Physica E* **23**, 171 (2004).
21. S. Mendach, O. Schumacher, Ch. Heyn, S. Schnüll, H. Welsch, and W. Hansen, "Preparation of curved two-dimensional electron systems in InGaAs/GaAs-microtubes," *Physica E* **23**, 274 (2004).
22. S. Mendach, O. Schumacher, H. Welsch, Ch. Heyn, W. Hansen, and M. Holz, "Evenly curved two-dimensional electron systems in rolled-up Hall bars," *Appl. Phys. Lett.* **88**, 212113 (2006).
23. N. Shaji, H. Qin, R. H. Blick, L. J. Klein, Ch. Deneke, and O. G. Schmidt, "Magnetotransport through two dimensional electron gas in a tubular geometry," *Appl. Phys. Lett.* **90**, 042101 (2007).

24. A. B. Vorob'ev, K.-J. Friedland, H. Kostial, *et al.* "Giant asymmetry of the longitudinal magnetoresistance in high-mobility two-dimensional electron gas on a cylindrical surface," *Phys. Rev. B* **75**, 205309 (2007).
25. M. V. Entin and L. I. Magarill, "Electrons in a twisted quantum wire," *Phys. Rev. B* **66**, 205308 (2002).
26. M. Trushin and J. Schliemann, "Spin dynamics in rolled-up two-dimensional electron gases," *New J. Phys.* **9**, 346 (2007)
27. V. Sverdlov, E. Ungersböck, and H. Kosina, "Theoretical electron mobility analysis in thin-body FETs: Dependence on substrate orientation and biaxial strain," *IEEE Trans. Nanotechnol.* **6**, 334 (2007).
28. Yu. S. Yukecheva, G. Mussler, V. Shushunova, *et al.*, "Modulation-doped Si_xGe_{1-x}/Si shells electrically isolated from conductive substrates," *Semicond. Sci. Technol.* **23**, 105007 (2008).
29. M. Grundmann, "Nanoscroll formation from strained layer heterostructures," *Appl. Phys. Lett.* **83**, 2444 (2003).
30. S. V. Golod, V. Ya. Prinz, P. Wägli, *et al.*, "Freestanding SiGe/Si/Cr and $SiGe/Si/Si_xN_y/Cr$ microtubes," *Appl. Phys. Lett.* **84**, 3391 (2004).
31. F. Schäffler, "High-mobility Si and Ge structures," *Semicond. Sci. Technol.* **12**, 1515 (1997).
32. M. H. White and J. R. Cricchi, "Characterization of thin oxide NMOS memory transistors," *IEEE Trans. Electron Dev.* **19**, 1280 (1972).
33. H. Lueth, *Surfaces and Interfaces of Solids*, Berlin: Springer, 1993.
34. C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys.* **55**, 645 (1983).
35. G. Etemadi and J. F. Palmier, "Effect of interface roughness on nonlinear vertical transport in GaAs/AlAs superlattices," *Solid State Commun.* **86**, 739 (1993).
36. S. Tsujino, C. Falub, E. Müller, M. Scheinert, L. Diehl, U. Gennser, and T. Fromherz, "Hall mobility of narrow $Si_{0.2}Ge_{0.8}$–Si quantum wells on $Si_{0.5}Ge_{0.5}$ relaxed buffer substrates," *Appl. Phys. Lett.* **84**, 2829 (2004).

Silicon Nanowire-Based Nonvolatile Memory Cells: Progress and Prospects

Qiliang Li, X. Zhu, Y. Yang, D. E. Ioannou
ECE Dept., George Mason University, Fairfax, VA 22030, USA

J. S. Suehle and C. A. Richter
Semiconductor Electronics Division, NIST, Gaithersburg, MD 20899, U.S.A.

1. Introduction

There is a widely recognized need for urgent progress in nonvolatile memory (NVM) technology,[1] driven by the exponential market growth of the ubiquitous portable and mobile electronics, such as cell phones, music players, USB and memory cards, *etc.* Commensurate with this demand, intensive research is carried out around the world to improve present technologies and invent new ones: for example, no fewer than eight (out of a total of 31) dedicated technical sessions at the IEDM 2007 meeting[2] and one[3] out of three plenary talks at the IEDM 2008 meeting[4] were dedicated on memory cell technology. To meet this demand and enable continued scaling (Moore's Law), NOR flash memory is been replaced (where possible) with the much denser NAND technology, but the industry consensus is that fundamentally new innovations must be introduced beyond the 45-nm lithography node generation.[3] One such innovation actively pursued by many research groups around the world is to design NVM cells based on silicon nanowires (SiNWs).[5]

There are several important reasons why SiNW NVM cells are expected to outperform a conventional (planar)[6] silicon NVM cell. First, because of the cylindrical symmetry, for the same value of the control gate voltage the electric field across the tunneling oxide is bigger than in comparable conventional cells, leading to either higher performance (faster program/erase or P/E), or same performance at a lower power operation. Second, the cylindrical symmetry is particularly suitable for scaling NVM cells, where as short a channel length as possible is desired without reducing the gate oxide thickness too much so that good retention is maintained. From the expression for the relevant "natural scaling length" λ_{CYL}:[7,8]

$$\lambda_{CYL} = t_{Si}\sqrt{\frac{2\varepsilon_{Si}\ln(1+2t_{OX}/t_{Si})+\varepsilon_{OX}}{16\varepsilon_{OX}}}$$

it is clearly seen that while λ_{CYL} (and thus the channel length) scales with t_{Si} (nanowire diameter), it only depends logarithmically on the tunneling oxide thickness t_{OX}.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

In this chapter, the most relevant results from the literature on SiNW NVM cells will first be reviewed, and our latest results will be presented and discussed. An attempt will then be made to project future progress and anticipate new innovations on SiNW NVM cells.

2. A brief review of nanowire-based NVM cells

The first nanowire non-volatile memory cells were actually not based on silicon nanowires, but used carbon nanotube (CNT) channels,[9] grown on a conducting Si wafer capped with quite thick (500 nm) thermally grown SiO_2 as the storage node, the wafer itself acting as the back gate. As the gate voltage was swept back and forth between +10 V and −10 V, a large hysteresis was evident in the $I_D(V_G)$ curves and a threshold gate voltage V_{TH} shift (*i.e.* a memory window) of more than 6 V was obtained. In addition, Fuhrer and co-workers predicted that memory cells with such narrow, high mobility nanotube field-effect transistors (FETs) should significantly outperform cells based on conventional (planar) FETs in charge detection sensitivity. Nanoscale memory cells with similar device structures were also reported by Radosavljevic *et al.*[10] based on ambipolar carbon nanotube FETs. They exploited the resulting current-voltage hysteresis to create a NVM cell with a single CNT FET that was stable at room temperature. Soon afterwards Choi *et al.* patented[11] a CNT-based nonvolatile memory cell structure with oxide-nitride-oxide (ONO) dielectric stack and explained its operation and advantages.[12] The ONO layer had high breakdown voltage, low defect density, and high charge retention capability and in their cells the thicknesses of the O/N/O layers were 7 nm, 7 nm, and 14 nm, respectively. The gate electrode was deposited over the ONO stack (top-gate) and was located within ("under-lapped") the predefined CNT channel. Device operation was tested from low temperature to room temperature. At 5 K, a clear $I_D(V_G)$ hysteresis was observed when the gate voltage was swept over 4 V, and the threshold voltage shift was about 2 V when the gate voltage was swept over 12 V. Since the diameter of the CNT was about 3 nm, a high electric field was produced around the surface of CNT. The calculated electric field for $V_G = 5$ V was found to be high enough to produce Fowler-Nordheim tunneling. These results strongly indicated that the CNT memory could be a candidate for ultrahigh-density flash memory. Ganguly *et al.*[13] fabricated carbon nanotube-based nonvolatile cells with charge storage in metal nanocrystals (NCs) and later analyzed the resulting enhanced electrostatics.[14] The devices had a large memory window with low operating voltage and single electron controlled drain current, and it was suggested that Coulomb blockade in the nanocrystals combined with single charge sensitivity of the nanotube FET can potentially enable multilevel memory operation. Ganguly *et al.*[13] noted an important difference between NVM cells with one-dimensional (1D) and two-dimensional (2D) channels and discrete charge storage: 1D channel current is mainly controlled by the maximum barrier in the channel and can be modulated by a single charged nanocrystal near the 1D channel (known as the "bottleneck" effect), whereas in the 2D channel case, the least resistive path controls the current. The bottleneck effect was very strong due to the combination of nanocrystals and CNT, and contributed greatly to achieving large memory window and single-electron sensitivity of the device. Their simulations demonstrated that the NCs were electrostatically more strongly coupled to the 2D gate electrode than to the 1D channel, even when the NCs were at a much closer proximity to the 1D channel, leading to efficient tunneling and low-voltage program operation. Under the retention condition, the NC-CNT devices had lower electric field across the tunneling oxide than that in the case of a 2D channel. This enhanced electric field difference with respect to program *vs.* retention operations produced larger ratio between program and retention times. Together with the large number of electrons stored per NC, this enhanced electrostatics can be utilized either to reduce the operating voltage or to reduce statistical fluctuation of the information storage. More recently, Cha *et al.* fabricated NVM cells with gallium nitride nanowire and top gate[15] and explained the operation mechanism. A 40 nm of top gate dielectric SiO_2 was used as the charge-trapping medium. A simple 2D simulation was performed, with the result that the electric field distribution was different between the center and the edge of the gate. While the electric field distribution under the center of the gate is uniform across the oxide layer, the field under the gate edge changes along the distance. In the write mode, charges in the gate are injected into the oxide with the help of the locally enhanced electric field around the edge of the nanowire, and they continue to flow toward the nanowire channel (by Ohmic conduction mechanisms) until they lose energy and/or are captured by traps in the low field regime (away from the nanowire corners). In the erase mode, the locally enhanced electric field moves the trapped charges back to the gate.

3. Silicon nanowire-based NVM

We have reported Si nanowire based NVM in 2007.[5] The devices were fabricated with Si nanowires grown from the Au catalyst on the predefined locations on the $SiO_2/Si_3N_4/SiO_2$ surface. A large memory window of 17 V was observed from the SiNW ONO cell when the back gate bias was swept over 20 V. Simple, reversible write, read and erase operations were demonstrated as well.

The active medium of these devices is a nanowire channel grown by the "bottom-up" method. In the last few years, SiNW memory cells fabricated by the "top-down" approach were also demonstrated. For example, in 2007 Suk *et al.*[16,17] from Samsung Electronics presented silicon-oxide-nitride-oxide-silicon (SONOS) gate-all-around (GAA) memory with twin Si nanowires fabricated by using a self-aligned damascene-gate process. The ONO gate dielectric stacks were deposited around the SiNWs. The electrical measurement results indicated that smaller diameter nanowires had faster program/erase speed due to the larger vertical electrical field at the interface between SiNW and the tunneling oxide. Meanwhile, the Singh *et al.*[18,19] research group in Singapore reported GAA MOSFET devices with small diameter twin SiNWs (vertically stacked), which were fabricated by using a fully CMOS-compatible alternating phase shift mask (alt-PSM) method. With this novel memory structure, Fu *et al.* demonstrated several other memory device variants.[20-23] Vertically stacked SiNWs have the advantage of a reduced footprint per NW channel. Compared with the conventional planar devices, these devices exhibited large drive current, excellent gate control and much-improved program/erase speed at relatively small voltages. These enhancements of device performance are attributed to the optimized electron energy distribution, the potential energy profile, and electric field along each layer surrounding the channel. Furthermore, these works have introduced high-κ dielectric stacks TaN/Al_2O_3/HfO_2/ SiO_2/Si (TAHOS)[22] or polycrystalline Si nanowire to implement charge trapping memory cells.[23]

Our group has been engaged on developing SiNW based NVM by using "bottom-up" self-assembled SiNW as channels and high- κ dielectric stacks for charge storage.[24-28] In the following sub-sections, we will describe the design and fabrication process of SiNW-based NVM cells. The electrical characterization and modeling of these memory cells will be demonstrated and discussed as well.

- *Experimental*

In our work, the SiNW-based NVM cells were fabricated by a "self-alignment" process. The essential steps are illustrated in Fig. 1. First, a layer of thermal SiO_2 was grown by dry oxidation on a silicon wafer, and the SiNWs (~ 20 nm of diameter, Fig. 2) were then grown from Au catalysts in pre-defined locations by chemical vapor deposition (CVD) – see Fig. 3(a) and (b). The SiNW growth was followed by thermal oxidation to grow the tunneling SiO_2 layer (~3.5 nm) surrounding the SiNW. Aluminum Schottky contact source/drain electrodes were patterned on the SiNW by lithography (Fig. 3(c)). The next process step was the ALD of high-κ HfO_2 as the charge trapping layer and ALD of Al_2O_3 or sputtered SiO_2 (for the control NVM cell H1) to form the blocking layer, which almost completely surrounds the SiNW channel, shown in Fig. 4(a).

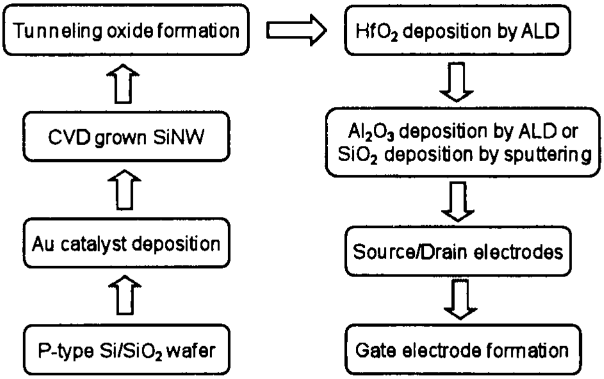


Figure 1. Main fabrication steps of the self-alignment process flow.

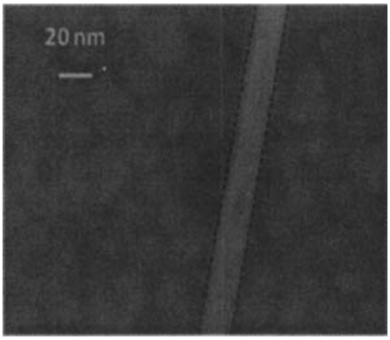


Figure 2. Scanning electron microscope (SEM) image of a single SiNW grown by chemical vapor deposition.

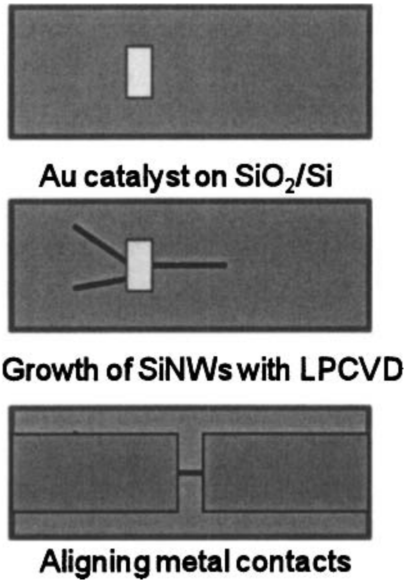


Figure 3. Illustration of the basic steps of the self-aligning process: (a) Au catalyst is patterned on SiO$_2$/Si substrate; (b) SiNWs are grown from Au catalyst by CVD; (c) top view of the contacts to a typical SiNW memory cell.

Finally, an Al gate electrode was defined and deposited. By optimizing the catalyst position and device structure, about 90% of the predefined locations form expected SiNW devices with one or more SiNW connected. A typical SiNW NVM cell with a 5 μm gate length is shown in Fig. 4(b). This *in-situ* fabrication process leads to clean interfaces with much better electronic properties compared to the commonly used alternative processes that depend on microfluidic or electrophoretic alignment.

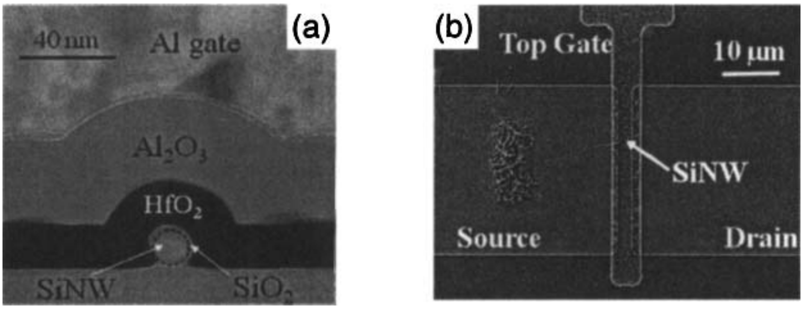


Figure 4. (a) Cross-sectional SEM micrograph of a typical cell. The gate dielectric stack almost completely surrounds the SiNW; (b) top view SEM micrograph.

- *Results and discussion*

Figure 5(a) shows the $I_D(V_{GS})$ characteristics of the initial and the programmed states of the memory cell (HfO_2/Al_2O_3 of 20/40 nm). The magnitudes of the memory windows for different dielectric thickness are listed in Fig. 5(b), which indicates that the memory window depends on both the dielectric thickness and program/erase voltage. As the thickness of the charge-trapping HfO_2 increases, the number of available trapping centers increases, leading to the larger observed memory window. As the P/E voltage increases, the electric field at the interface between SiNW and tunneling oxide increases, resulting in more electrons tunneling to the charge storage layer. Table 1 lists the thickness of HfO_2 and Al_2O_3 (or SiO_2) layers for the three cells (H1–H3) that will be presented and discussed here. Figure 6 shows the programming and erasing characteristics of cells H1-H3 at different P/E gate bias levels with source and drain electrodes kept grounded. It should be noted that cell H3, which has the thinnest charge-trapping layer (20 nm) and blocking oxide (20 nm), shows the fastest P/E speed.

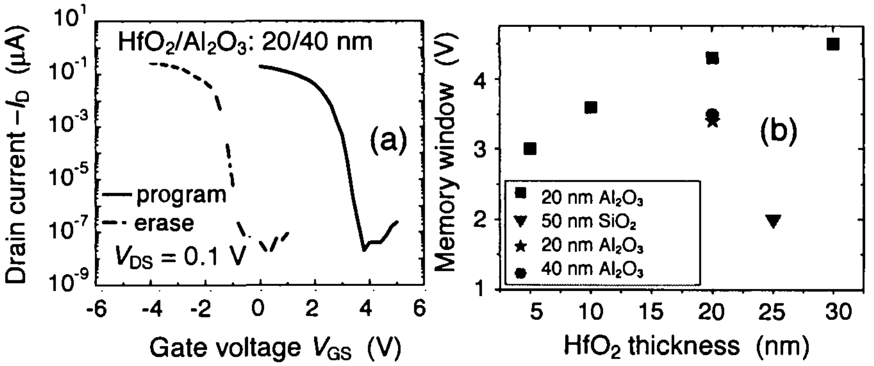


Figure 5. (a) Current–voltage characteristics of the memory cell (HfO_2/Al_2O_3 of 20/40 nm) in initial and programmed states; (b) memory windows for cells with different dielectric thickness and P/E bias conditions (squares denote V_P = 14 V for 3 seconds, other symbols denote V_P = 10 V for 2 seconds).

Device	Charge Trapping Layer	Blocking Oxide
NVM Cell H1	25 nm HfO_2 by ALD	50 nm of SiO_2 by Sputtering
NVM Cell H2	20 nm HfO_2 by ALD	40 nm of Al_2O_3 by ALD
NVM Cell H3	20 nm HfO_2 by ALD	20 nm of Al_2O_3 by ALD

Table 1. SiNW NVM cells H1-H3 indicating trapping and blocking layer thicknesses.

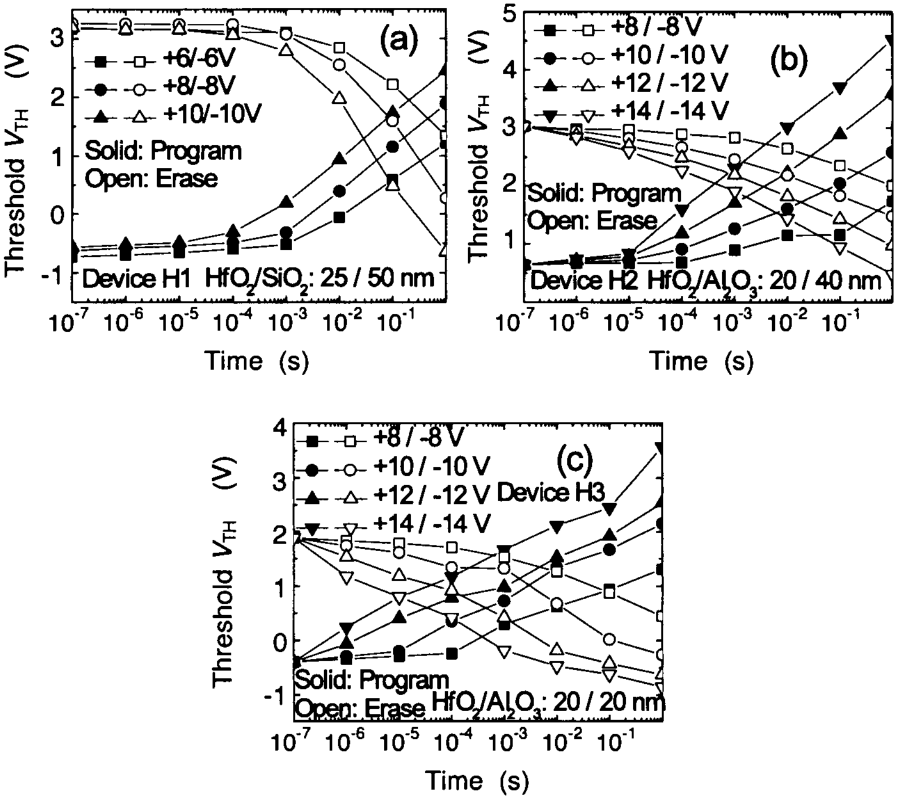


Figure 6. Program/erase characteristics for cells H1–H3 from Table 1: (a) H1; (b) H2; and (c) H3.

Figure 7 shows the numerical simulated profiles of the electric field and the electrostatic potential through the structure of cells H1–H3 under a gate voltage of +10 V and all other terminals grounded. It is clear from Fig. 7 that the electric field intensity at the interface of SiNW and tunneling oxide in H3 is much larger than the other two devices, which is consistent with the results shown in Fig. 6.

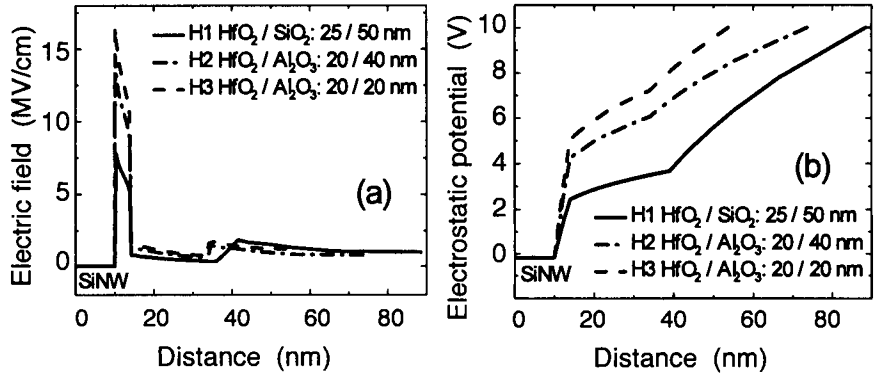


Figure 7. Simulation profiles of (a) the electric field and (b) the electrostatic potential through the structure of cells H1–H3.

Figure 8(a) shows the cell endurance where an upward shift of the threshold voltage (V_{TH}) is observed for both programmed and erased states after a number of P/E cycles. The cells maintain the same memory window (no noticeable degradation) after 10^4 P/E cycles. The program and erase pulses were +8 V for 1 ms and –8 V for 10 ms, respectively. The upward shift of V_{TH} may be due to accumulation of electrons trapped in residual deep traps in the HfO_2 layer. Figure 8(b) shows the retention characteristics at room temperature, from which the 10-year retention threshold voltages (estimated by linear extrapolation) show a trapped charge loss of ~50%, ~11% and ~9.1% for H1, H2, and H3, respectively. The observed large charge loss of cell H1 most likely arises from the poor quality of the sputtered SiO_2 as the blocking oxide, whereas the superior charge retention of cells H2 and H3 results from the high quality of the ALD-deposited Al_2O_3 blocking layer.

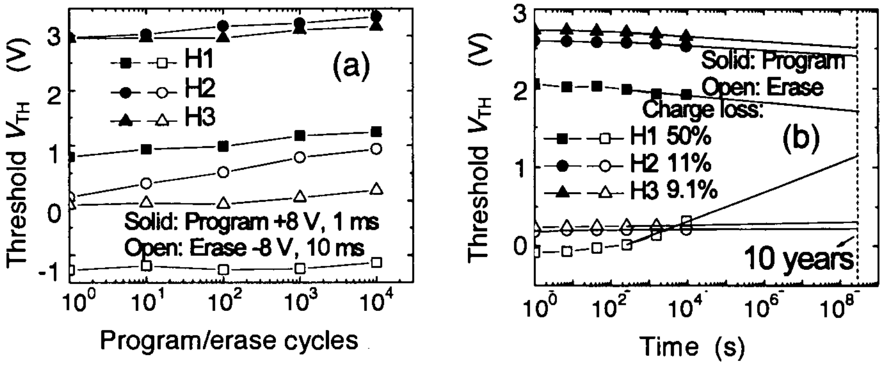


Figure 8. (a) Endurance characteristics of cells H1–H3; (b) retention characteristics of cells H1–H3 at room temperature.

4. Conclusions and future prospects

The semiconductor memory industry is experiencing major technological changes with the introduction of three-dimensional multi-gate structures and charge trapping. Nanowires have attracted much attention due to their intrinsic scalability and natural cylindrical structure. The combination of nanowire channel and high-κ gate dielectric stacks is an very attractive strategy to implement three-dimensional charge-trapping memory with high performance at low cost, which has been continuously demonstrated by the recent progress in research of nanowire based charge trapping memory cells. Several technical issues of nanowire high-κ dielectric memory are targeted and under intensive study, including how to improve the quality of nanowire/dielectric interface, how to reduce the cost, and how to integrate it for high-speed programming for application in standalone memories and embedded systems.

References

1. S. K. Lai, "Flash memories: Successes and challenges," *IBM J. Res. Develop.* **52**, 529 (2008).
2. See *Tech. Dig. IEDM* (2007).
3. S. K. Lai, "Non-volatile memory technologies: The quest for ever lower cost," *Tech. Dig. IEDM* (2008), p. 11.
4. See *Tech. Dig. IEDM* (2008).
5. Q. Li, X. Zhu, H. D. Xiong, *et al.*, "Silicon nanowire on oxide/nitride/oxide for memory application," *Nanotechnology* **18**, 235204 (2007).
6. P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells – An overview," *Proc. IEEE* **85**, 1248 (1997).
7. C. P. Auth and J. D. Palmer, "Scaling theory for cylindrical, fully-depleted, surrounding gate MOSFETs," *IEEE Electron Dev. Lett.* **18**, 74 (1997).
8. J. -P. Colinge, "Multiple-gate SOI MOSFETs," *Solid State Electronics* **48**, 897 (2004).
9. M. S. Fuhrer, B. M. Kim, T. Durkop, and T. Brintlinger, "High mobility nanotube transistor memory," *Nano Lett.* **2**, 755 (2002).
10. M. Radosavljevic, M. Freitag, K. V. Thadani, and A. T. Johnson, "Nonvolatile molecular memory elements based on ambipolar nanotube field effect transistors," *Nano Lett.* **2**, 761 (2002).
11. W. B. Choi, I. Yoo, and J. Chu, Korean Patent No. KR 20020007709.
12. W. B. Choi, S. Chae, E. Bae, *et al.*, "Carbon-nanotube-based nonvolatile memory with oxide-nitride-oxide film and nanoscale channel," *Appl. Phys. Lett.* **82**, 275 (2003).
13. U. Ganguly, E. C. Kan, and Y. Zhang, "Carbon nanotube-based nonvolatile memory with charge storage in metal nanocrystals," *Appl. Phys. Lett.* **87**, 043108 (2005).

14. U. Ganguly, C. Lee, T.-H. Hou, and E. C. Kan, "Enhanced electrostatics for low-voltage operations in nanocrystal based nanotube/nanowire memories," *IEEE Trans. Nanotechnology* **6**, 22 (2007).
15. H.-Y. Cha, H. Wu, S. Chae, and M. G. Spencer, "Gallium nitride nanowire nonvolatile memory device," *J. Appl. Phys.* **100**, 024307 (2006).
16. S. D. Suk, K. H. Yeo, K. H. Cho, *et al.*, "Gate-all-around twin silicon nanowire SONOS memory," *Tech. Dig. VLSI Symp.* (2007), p. 142.
17. S. D. Suk, S.-Y. Lee, S.-M. Kim, *et al.*, "High performance 5 nm radius twin silicon nanowire MOSFET (TSNWFET): Fabrication on bulk Si wafer, characteristics, and reliability," *Tech. Dig. IEDM* (2005), p. 717.
18. N. Singh, A. Agarwal, L. K. Bera, *et al.*, "High-performance fully depleted silicon nanowire (diameter $\leq$ 5 nm) gate-all-around CMOS devices," *IEEE Electron Dev. Lett.* **27**, 383 (2006).
19. N. Singh, F. Y. Lim, W. W. Fang, *et al.*, "Ultra-nanowire silicon nanowire gate-all-around CMOS devices: Impact of diameter, channel-orientation and low temperature on device performance," *Tech. Dig. IEDM* (2006), p. 547.
20. J. Fu, K. D. Buddharaju, S. H. G. Teo, *et al.*, "Trap layer engineered gate-all-around vertically stacked twin Si-nanowire nonvolatile memory," *Tech. Dig. IEDM* (2007), p. 79.
21. J. Fu, N. Singh, K. D. Buddharaju, *et al.*, "Si-nanowire based gate-all-around nonvolatile SONOS memory cell," *IEEE Electron Dev. Lett.* **29**, 518 (2008).
22. J. Fu, N. Singh, B. Yang, C. X. Zhu, G. Q. Lo, and D. L. Kwong, "Si-nanowire TAHOS (TaN/Al_2O_3/HfO_2/SiO_2/Si) nonvolatile memory cell," *Proc. ESSDERC* (2008), p. 115.
23. J. Fu, Y. Jiang, N. Singh, C. X. Zhu, G. Q. Lo, and D. L. Kwong, "Polycrystalline Si nanowire SONOS memory cell fabrication on a gate-all-around (GAA) channel architecture," *IEEE Electron Dev. Lett.* **30**, 246 (2009).
24. X. Zhu, D. Gu, Q. Li, D. E. Ioannou, H. Baumgart, J. S Suehle, and C. A. Richter, "Silicon nanowire NVM with high-κ gate dielectric stack," *Microelectronic Eng.* **86**, 1957 (2009).
25. X. Zhu, Y. Yang, Q. Li, D. E. Ioannou, J. S. Suehle, and C. A. Richter, "Silicon nanowire NVM cell using high-κ dielectric charge storage layer," *Microelectronic Eng.* **85**, 2403 (2008).
26. X. Zhu, Y. Yang, Q. Li, D. E. Ioannou, J. S. Suehle, and C. A. Richter, "High performance silicon nanowire field effect transistor and application to non-volatile memory," presented at Washington Acad. Sci. Conf. (2008).
27. X. Zhu, Q. Li, Y. Yang, *et al.*, "Fabrication and characterization of silicon nanowire memory with hafnium oxide charge storage layer," presented at *MRS Spring Meeting* (2008).
28. X. Zhu, Q. Li, D. E. Ioannou, *et al.*, "Silicon nanowire memory application using hafnium oxide charge storage layer," *Proc. Intern. Semicond. Dev. Res. Symp.* (2007), p. 22.

Prospects and Challenges of Next-Generation Flash Memory Devices

Jang-Sik Lee
School of Advanced Materials Engineering, Kookmin University
Seoul 136-702, Korea

1. Introduction

We can divide the history of memory devices in terms of their applications. From the latter half of the 1980's, personal computers were the major application area, and in that regime, DRAM was the dominant memory device. However, since latter half of the 1990's, the use of mobile electronic devices, such as MP3 players, digital still cameras, personal digital assistants, cellular phones, *etc.*, has increased greatly, making flash memory devices the most widely used form of memory. The rapid progress in this field has been quite beyond expectations.[1-5] However, we now face very difficult challenges inherent in the development of NAND flash memory devices with a 20 nm design rule. Further, scaling predicts that memory devices with feature sizes down to 10 nm should be developed in just a few years. It is not clear how this will be achieved. For now, lithography advances have made it possible to use optical lithography tools to define 30 nm features and thanks to this, we can develop 10+ Gb memory devices without resorting to nonoptical lithography, such as EUV, electron-beam, or x-ray.

In addition to the rapidly advancing technology, the sales of flash memory devices are also very encouraging. Since 2005, NAND flash revenues have surpassed those for NOR flash and NAND will continue to be a leading player in overall flash market. Also, it is expected that NAND will overtake DRAM by 2011. Cost per bit has been declined dramatically and the NAND market has been expanding quickly. The increase of the NAND market share and decrease of cost have been driven by aggressive scaling and novel technology, such as multi-level data storage, *etc.* These trends will undoubtedly open up new application areas. For example, it is now possible to replace the hard disk drive by a solid-state drive (SSD) made using flash memory devices. In the near future, we can expect server machines to be equipped with SSDs with massive memory capacity.

Not only has NAND memory outpaced other Si devices in terms of rapid progress, but in 2001 NAND became the first mass-produced device using real nanotechnology. This trend is still ongoing; with volume production of 32 Gb NAND flash fabricated with a 30 nm design rule to begin this year. Accordingly, this is the right moment to look at the challenges we face, as well as possible solutions and innovations for further device scaling, and future directions for flash memory development.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

2. Scaling challenges

As device scaling continues, we face numerous scaling challenges. First, the tunneling oxide thickness should be reduced, but this is very difficult because of charge retention and stress-induced leakage current issues. Also, there are serious problems associated with higher cell packing density, such as increasing cell-to-cell interference. As the dimensions of the floating gate become smaller, the total number of stored electrons in floating gate is reduced, resulting in very tight electron loss margins. Also, as the dimensions of the NAND flash decrease, the parasitic capacitance increases, leading to a smaller coupling ratio – see Fig. 1(a).[4] Another serious problem is cell-to-cell interference, which increases as the cells become smaller and closer together, resulting in the cell V_T shifts and decreasing memory window margins. With decreasing area of the interpoly dielectric (IPD) layer, typically oxide/nitride/oxide (ONO), the capacitance also decreases. Correspondingly, the total number of stored charge carriers decreases, resulting in reduced charge loss tolerance, which is further exacerbated by tighter design rule. At 30 nm process technology, a flash memory cell transistor is expected to store only around 100 electrons, see Fig. 1(b).[4] This means the charge loss tolerance will be around 10 electrons – a very difficult challenge to overcome.

However, many innovations have been introduced during the last several years to overcome these scaling challenges. To reduce the cell-to-cell interference, new device schemes were introduced, *e.g.* charge-trap flash memory devices. Further downscaling has been achieved by introducing double patterning in the existing ArF lithography tools. To increase the memory density without cell size reduction, three-dimensional (3D) stacked integration and multi-level cell schemes were introduced. High-κ IPD floating-gate cell structures were developed to combat cell-to-cell interference. To use the chip area effectively, the length of the NAND cell string has been continuously increased, most recently from a 16-cell string to a 64-cell string, to keep the overhead low as the number of strings has been increasing with each technology node.

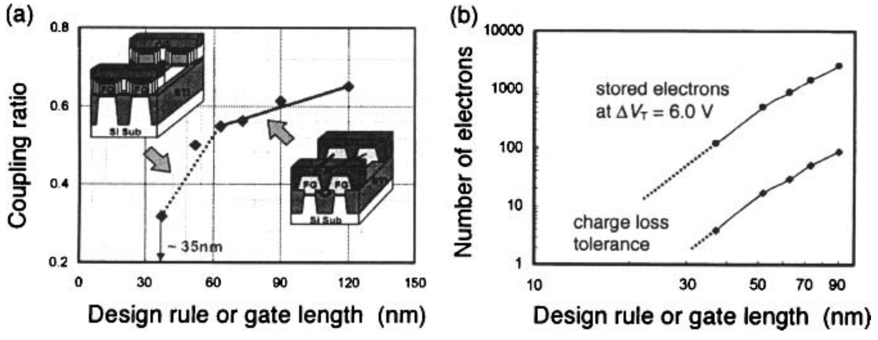


Figure 1. (a) NAND coupling ratio changes according to the design rule; (b) number of stored electrons in unit flash memory cell and charge loss tolerance as a function of design rule. [Adapted from Ref. 4, copyright 2005, IEEE].

3. Technological innovations

We continuously face limitations in the field of microelectronics. These limitations can be either physical or electrical. Even the ITRS Roadmap predicts that flash scaling will reach its limit within the next several years. So what will happen next? The volume production of 32 Gb NAND flash has already started at Samsung, Toshiba, and possibly other semiconductor companies. Within a couple of years, we will need to develop higher density memory devices to maintain the scaling trends. We can continue to scale down the flash memory devices by using the following strategies. First, we can continue to reduce the cell size. It appears that 193 nm immersion lithography, resolution enhancement techniques (RET) and double patterning technology (DPT) can be used to produce NAND flash memory devices down to the 20 nm design rule without resorting to nonoptical lithography. Another option is to use multi-level data storage. Multi-level cells (MLCs) are already a confirmed technology at the 2 bits/cell level, but nowadays super-MLCs are actively developed, aiming for 3 bits/cell (8 data levels) and 4 bits/cell (16 data levels). These super-MLCs can be used to increase memory density without reducing the cell size for another a couple of generations. Finally, we have 3D stacked structures, again to increase the memory density without reducing the cell size, which involve 3D integration and 3D packaging. All of these approaches are illustrated in more detail below.

- *Novel lithographic technologies*

First, let us look at the continuous scaling by using optical lithography. Claims that optical lithography is reaching its limits and that next-generation lithography will soon be required have been with us for many years. Typically, the proposed next-generation techniques have referred to EUV, electron-beam, or x-ray lithography approaches. Nonetheless, the longevity of optical lithography has surpassed such expectations and it remains the dominant technology today, the only one capable of confronting the device scaling challenges in volume production. The survival of optical lithography is based on continuous technological progress. A particularly major advance was the introduction of immersion lithography. According to the famous Rayleigh criterion, the resolution limit is given by $k_1\lambda/NA$, where λ is the wavelength and NA the numerical aperture. Since the wavelength of the optical source is limited, definition of ever finer patterns required either a decrease in k_1 or an increase in NA. As long as the lithographic exposure was done in air, the NA value was limited by the size of the optics. In immersion lithography, by using water as contact media between the objective lens and the resist, the NA value could be significantly increased. Today, most semiconductor companies have adopted immersion lithography tools to obtain finer patterns.

Another important advance was the double patterning technique (DPT). Several different DPT approaches, such as double exposure, double pattering, and spacer double patterning, are illustrated in Fig. 2.[6] These approaches are slightly different, but lead to similar results.

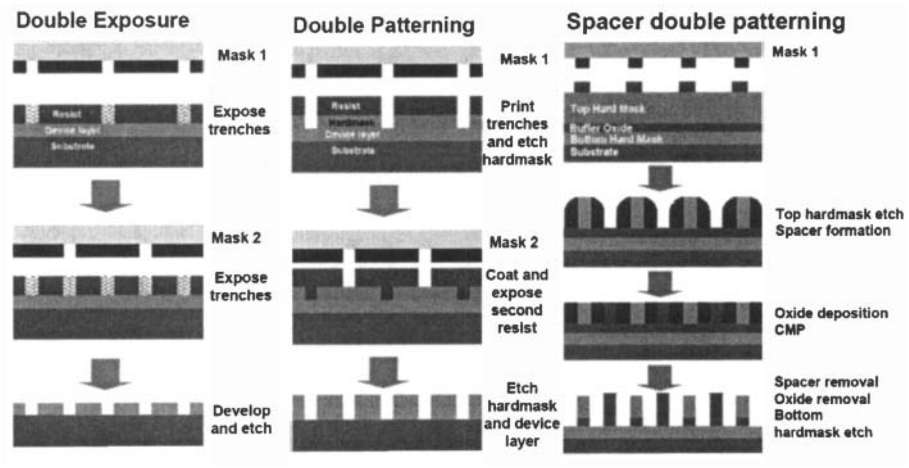


Figure 2. Schematic illustration of the double patterning techniques (DPT) that can be done by using double exposure, double patterning, and spacers. Adapted from Ref. 6.

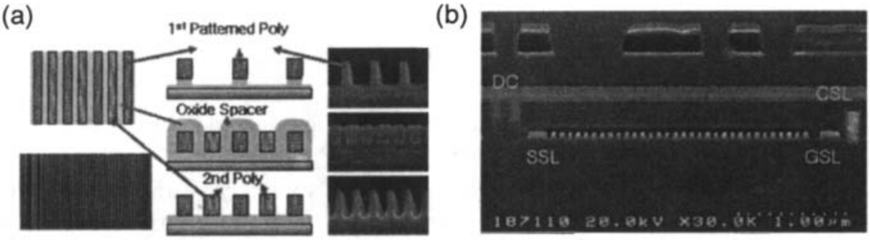


Figure 3. (a) Schematic illustration of self-aligned double patterning technique (SADPT) to define the poly-layer and (b) fully-processed NAND flash cell string using SADPT. [Adapted from Ref. 7, copyright 2007, IEEE].

Figure 3 shows the first application of DPT to multi-gigabit NAND flash.[7] The first patterning was done using conventional 193 nm immersion lithography. Then, an oxide spacer was deposited on the patterned poly with a careful thickness control. A second poly layer was deposited in-between the oxide spacers. Finally, after removing the oxide spacer, well-aligned poly patterns of ~38 nm half pitch could be obtained, as shown in Fig. 3(a). This technique can used to define critical layers, *e.g.* active, poly-gate, bit line, *etc.* Figure 3(b) shows a fully processed NAND string by using DPT. This work was first presented by Samsung in early 2007 and was later employed to develop the first 64 Gb density NAND flash in November of 2007.[7] Thus, despite optical lithography limitations, it appears that

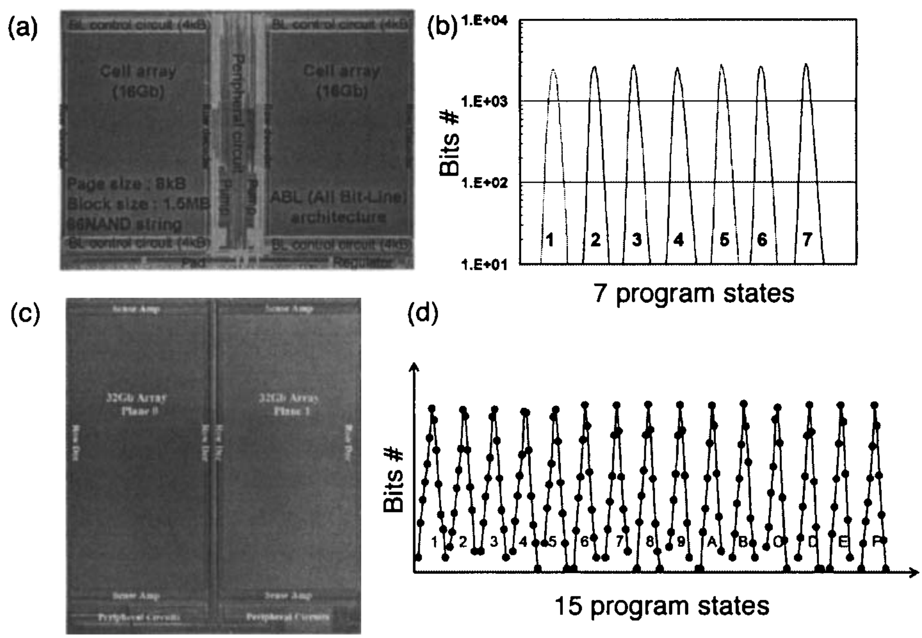


Figure 4. (a) Micrograph of 32 Gb NAND flash memory device and (b) cell V_T distribution with 7 programmed states (3 bits/cell). (c) Micrograph of 64 Gb NAND flash memory device and (d) cell V_T distribution with 15 programmed states (4 bit/cell). [Adapted from Refs. 8 and 9, copyright 2009, IEEE].

by further refining RET and using DPT, immersion lithography tools will be capable of patterning feature sizes as small as even 20 nm.

- *Super multi-level cells*

Basically most memory devices store 0 and 1 binary information by using program and erase operations. In flash memory, this type of storage is known as a single-level cell (1 bit/cell). However, it is possible to divide the program levels further by further refining the programmed states.

Thus, it is possible to store more information in unit cell, for example, 0, 1, 2, and 3, resulting in a 2 bits/cell MLC. By utilizing this method we can increase the density of memory device by a factor of two without reducing the cell size. Recently, further innovation in this field made it possible to store more than 2 bits per unit memory cell, for example, 3 bits/cell (8 data levels) and 4 bits/cell (16 data levels).[8,9] These types of cells are known as super-MLCs and this technique is being intensively explored, so we can increase the memory density without reducing the cell size for another a couple of generations. Figure 4 shows examples of NAND flash using 3 bit/cell and 4 bit/cell operations.

Figures 4(a) and 4(c) show the die micrographs of 32 Gb 3 bits/cell NAND flash and 64 Gb 4 bits/cell NAND flash, respectively. Figures 4(b) and 4(d) show the corresponding programmed cell distributions having 7 and 15 programmed states.[8,9] By using the super-MLC technique, it is possible to increase the memory density, but since the number of data levels cannot be increased indefinitely, 3D structures will provide another option to increase the memory density.

- *3D structures*

In the case of 3D structures, there are 2 different approaches: at the device integration level and at the packaging level. Let us consider 3D integration level first. Figure 5 shows the 3D stacked NAND cell string.[10,11] This was achieved by the selective epitaxial growth of single-crystal silicon on fully processed first memory layer. Then, the same process was done to fabricate the second memory layer. Since conventional process technology is used for both layers and common contact and source line can be used between two layers, the approach appears promising for device operation. However, the alignment of the 1st and 2nd layers is quite challenging and the electrical properties of epitaxially grown Si may be inferior to bulk Si, so device performance of the 1st and 2nd memory layers can be different. An alternative approach to device-level integration that is currently being explored is vertical NAND string integration.[12]

Another way to increase the memory density is chip-level multi-stacking. There are several ways to do this. For example, different processed device dies, for example SRAM, NOR, and NAND flash, can be integrated on a single package, making it possible to take advantage of each technology's unique properties. Also, 3D chip stacking, *e.g.* multiple NAND chips on single package, can also be employed to greatly increase the memory density. Recently, Toshiba showed that 128 GB SSD can be achieved with an area of 252 mm^2 by using stacked NAND flash memory chips as shown in Fig. 6.[13] This is a very good example for next-generation manufacturing of high-density NAND flash memory devices.

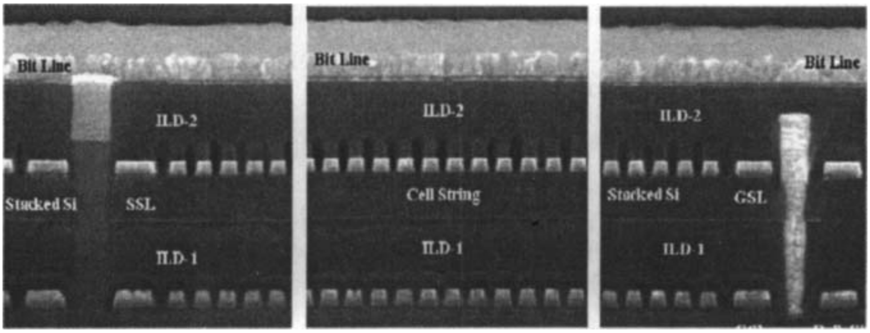


Figure 5. Scanning electron microscopy image of 3D stacked NAND flash string. [Adapted from Ref. 10, copyright 2006, IEEE].

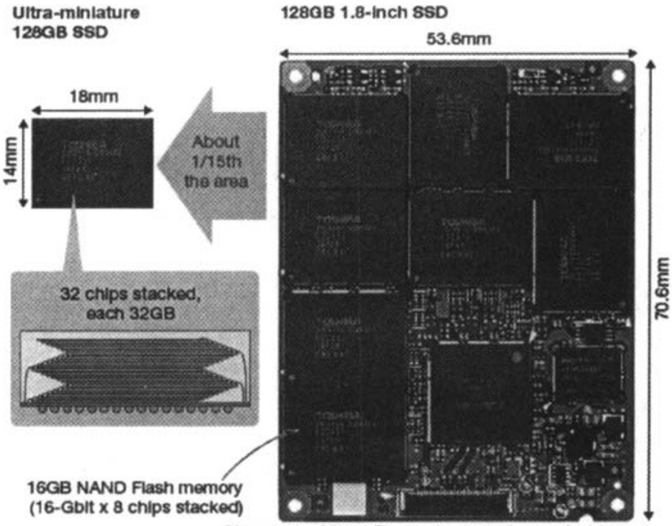


Diagram by *Nikkei Electronics* based on material courtesy Toshiba

Figure 6. 128 GB-density SSD produced by multiple stacking of NAND flash chips (left) and conventional 1.8-inch SSD (right). By using stacked structures the chip area can be effectively decreased. [Reproduced with permission from Nikkei Electronics Asia, *http://www.nikkeibp.com/neasia*].

4. Conclusions

Because of its impact on our daily life, flash memory has become a technology driver in the semiconductor industry. Flash memory development has been based on the continuous innovation in semiconductor fabrication and technology. By combining such advances as immersion lithography, double patterning, multi-level data storage, 3D stacked structures, *etc.* we can maintain the rapid NAND flash development towards ultra-high memory density. This trend will definitely open up new application areas as well as new markets.

Acknowledgments

I am grateful to Dr. Ki-Tae Park of Samsung Electronics for helpful discussions. This work was supported by the Center for Materials and Processes of Self-Assembly under the auspices of the ERC program of MEST/KOSEF (R11-2005-048-00000-0); "SystemIC2010" project of Korea Ministry of Knowledge Economy; Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (R01-2008-000-11994-0); Korea Research Foundation (KRF) Grant funded by the Korea Government (MEST) (KRF-2008-313-D00597); and the G.R.O.W. project of World Gold Council (RP05-08).

References

1. P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells – an overview," *Proc. IEEE* **85**, 1248 (1997).
2. R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to flash memory," *Proc. IEEE* **91**, 489 (2003).
3. C.-G. Hwang, "Nanotechnology enables a new memory growth model," *Proc. IEEE* **91**, 1765 (2003).
4. K. Kim, "Technology for sub-50 nm DRAM and NAND flash manufacturing," *Tech. Digest IEDM* (2005), pp. 323–326.
5. K. Kim and S. Y. Lee, "Memory technology in the future," *Microelectronic Eng.* **84**, 1976 (2007).
6. International Technology Roadmap for Semiconductors (ITRS) 2007 edition, Lithography section.
7. D. Kwak, J. Park, K. Kim, *et al.*, "Integration technology of 30 nm generation multi-level NAND flash for 64 Gb NAND flash memory," *IEEE Symp. VLSI Technol.* (2007), pp. 12-13.
8. T. Futatsuyama *et al.*, "A 113 mm^2 32Gb 3b/cell NAND flash memory," *Digest IEEE Intern. Solid-State Circ. Conf.* (2009), pp. 242–243.
9. C. Trinh *et al.*, "A 5.6 MB/s 64Gb 4b/Cell NAND flash memory in 43 nm CMOS," *Digest IEEE Intern. Solid-State Circ. Conf.* (2009), pp. 246–247.
10. S.-M. Jung, J. Jang, W. Cho, *et al.*, "Three dimensionally stacked NAND flash memory technology using stacking single crystal Si layers on ILD and TANOS structure for beyond 30 nm node," *Tech. Digest IEDM* (2006), art. 4154321.
11. K.-T. Park, D. Kim, S. Hwang, *et al.*, "A 45 nm 4 Gb 3D double-stacked multi-level NAND flash memory with shared bit line structure," *Digest IEEE Intern. Solid-State Circ. Conf.* (2008), pp. 510–511.
12. H. Tanaka, M. Kido, K. Yahashi, *et al.*, "Bit cost scalable technology with punch and plug process for ultra high density flash memory," *IEEE Symp. VLSI Technol.* (2007), pp. 14–15.
13. Nikkei Electronic Asia Magazine, June 2009, *http://techon.nikkeibp.co.jp/article/HONSHI/20090528/170917/*.

Chalcogenide Glassy Semiconductors – Could They Replace Silicon in Memory Devices?

K. D. Tsendin
Ioffe Physical-Technical Institute of Russian Academy of Sciences
St. Petersburg, Russia

1. Introduction

The phase-change memory in the chalcogenide glassy semiconductors based on a "crystal-glass" phase transition has been known since the mid-1960's.[1] The discovery of this phenomenon was not accidental, but was based on an understanding of the semiconducting properties of chalcogenide glassy semiconductors (CGS), which were discovered at Ioffe institute in 1955.[2]

Chalcogenide glassy semiconductors are semiconducting glasses incorporating elements of the VI group of the periodic table, such as sulfur, selenium, and tellurium. The ability of Ge-Te glasses to undergo fast crystalline-amorphous phase transition induced by a laser beam is being used in rewritable optical media, such as read/write CDs and DVDs. Moreover, a crystalline-amorphous phase transition may be produced by Joule heating due to an electric current. Accordingly, switching from a high (amorphous) to a low (crystalline) resistance state can be therefore used as a solid-state electronic memory, the more so as the resistance may be changed by many orders of magnitude.

The first generation of memory cells based on current-induced "crystal-glass" phase transitions was fabricated in the late 1970's. Unfortunately, these first-generation memory cells were prone to frequent data loss and hence unsuccessful. This failure can be attributed to the lacking understanding of the microscopic physical mechanisms underpinning the transition. Today, major semiconductor companies like Intel, Samsung and others are working on flash-type memory devices based on chalcogenide phase-change memory (PCM) cells. Such modern PCM cells are considered one of the promising directions for future nonvolatile memories.

Modern PCM cell operation is based on phase-change properties of a CGS, typically $Ge_2Sb_2Te_5$ (GST). Such GST-based PCM cells have demonstrated good cycling endurance of 10^{12} cycles, making a technology known from the mid-1960's relevant to nonvolatile memories of the 21st century. This chapter will examine the recent developments in microelectronics that have led scientists and technologists to consider the chalcogenide PCM cells as a possible alternative to the silicon-based flash memory devices.

First, a brief review of CGS switching and memory effects will be presented. Then, the properties of "on" and "off" states will be considered in more detail,

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

specially emphasizing the nonlinearity of current-voltage characteristics in the "off" state. We will show that information recording based on the crystal-glass phase transition induced by an electric field pulse in CGS has the peculiar property that the memory state arises not from a semiconducting but from a metallic state. This metallic state appears due to switching effects in thin films that have strongly nonlinear current-voltage characteristics.

2. Switching and memory effects

Generally, memory effects in solid-state materials arise from some switching effect.[1,3] A schematic diagram of the switching responsible for the PCM effect in a CGS film is sketched in Fig. 1. Each point of the high-resistance branch ("off" state) in Fig. 1(a),(b) corresponds to an applied voltage $V \leq V_T$, where V_T is a threshold voltage. If the applied voltage exceeds V_T, the material switches from the "off" state to the low-resistance "on" state after some delay t_D. The switching time t_{SW} of the transition is very small: $t_{SW} << t_D$. Furthermore, the delay t_D depends exponentially on the $(V - V_T)$ voltage difference and on the film thickness L. In a large device, the "on" state consists of a narrow high-current-density filament. If the material inside this filament does not crystallize for whatever reason – *e. g.*, CGS material with a weak tendency to crystallize, insufficiently high temperature in the filament, relatively small time t_{ON} in the "on" state, *etc.* – then for currents less than some on-state holding current I_H, the device will reversibly return to the "off" state. Thus, one has a reversible switching effect, illustrated in Fig. 1(a).

Conversely, if the material inside the high-current filament does crystallize because of an appropriately chosen CGS material, a sufficiently high filament temperature and a long enough t_{ON}, the device will not return to the "off" state, as shown in Fig. 1(b). After some set t_{SET}, the material in the filament will crystallize and remain crystalline form without any applied voltage, leading to the phase-change memory effect (PCM).

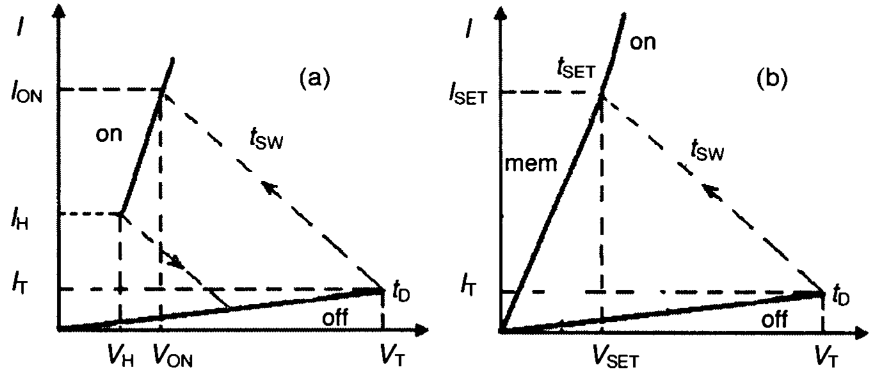


Figure 1. Schematic views of reversible switching (a) and PCM (b) current-voltage characteristics (see text).

Schematic diagrams of first-generation and modern CGS memory cells are presented in Fig. 2, while the current-voltage characteristic are presented in Fig. 3.

Let us consider in details the amorphous to crystalline transition (set process) and crystalline to amorphous transition (reset process). To obtain the set (crystalline) state one has to use current and voltage from the set region of Fig. 3 in order to reach the temperature of crystallization T_C, which is of the same order as the softening temperature T_G. Moreover, the crystallizing heat-pulse duration must be rather long for the crystallization to be completed. For the reset process (amorphization), one has to use current and voltage from the reset region of Fig. 3,

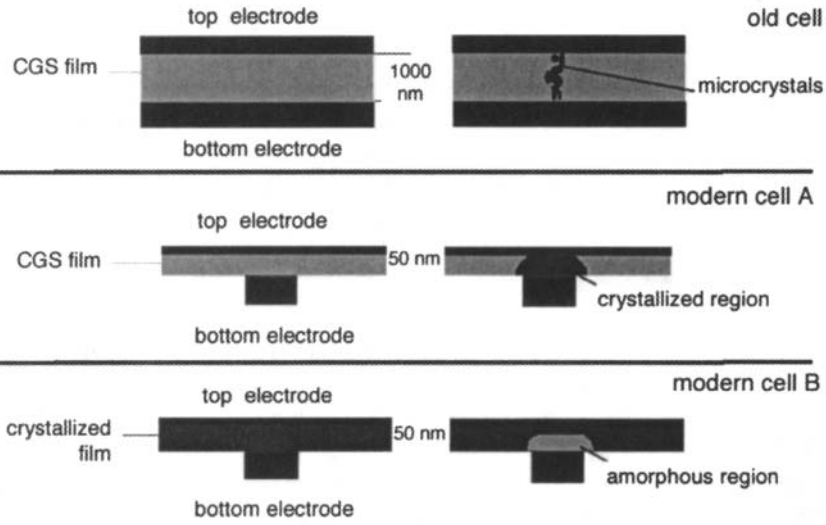


Figure 2. Schematic structures of the first-generation cell and two types of modern cells. The initial layer of the modern cells may be in the amorphous state (A) or in the crystalline state (B).

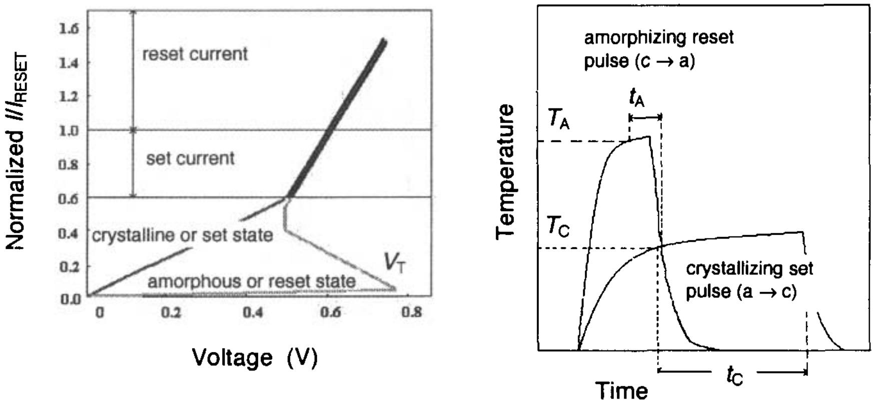


Figure 3. Current–voltage characteristic of a PCM type A cell of Fig. 2, fabricated from the popular $Ge_2Sb_2Te_5$ (GST) material.

with the material heated to the amorphization temperature T_A, which is of the order of melting temperature T_M. The reset pulse must be short with a rapid cooling slope in order to retain the amorphous state.

3. Properties of the "off", "on", and memory states

In the "off" state, the CGS are semiconducting glasses: $\sigma = \sigma_0 \exp(-E^*/kT)$, with an effective activation energy $E^* = E - f(V/V_0)$ that depends on the electric field. In the "on" state, E^* is approximately equal to zero, so one has the phase transition from semiconductor to metal conductivity, where $\sigma = \sigma_0$. The crystallization or memory set process changes the conductivity to yet another value, $\sigma_0 \rightarrow \sigma_C$. In other words, the memory state has a different conductivity from the "on" state. In a strong field of $\sim 10^5$ V/cm, the temperature T_{TR} of the phase transition from semiconductor to metal conductivity falls in the 450–600 K range for such compounds as $Si_{12}Te_{48}As_{30}Ge_{10}$ and As_2Se_3-$4As_2Te_3$. Table 1 summarizes the most important optical and electrical properties of the $Ge_2Sb_2Te_5$ (GST), which is currently considered the most promising CGS material for memory applications.

4. Comparison of Si-based and chalcogenide-based memory cells

The new generation PCM cells are variously known as phase-change random access memory (PRAM), Ovonic unified memory (OUM, to emphasize the contribution of S. Ovshinsky), and chalcogenide RAM (C-RAM). Figure 4 shows the modern PCM cell of the B type (see Fig. 2). A comparison of first-generation and modern PCM cells from the viewpoint of endurance is shown in Table 2. For the first-generation cell compounds the T_{TR} of the semiconductor to metal phase transition ($E^* = 0$) is shown, whereas the exact value for GST is not available.

	E_G (eV)	E_μ	E_n	$E = E_n + E_\mu$	n (cm^{-3})	μ cm^2/V·s	ρ (Ω cm) 300 K
amorphous CGS	0.74	0.25	0.2	0.45	$10^{17\text{-}18}$	10^{-2}	10^4
crystalline cubic (amorph → cubic at 440 K)	0.5	0	0.14	0.14	10^{16}	10^2	10^0
crystalline semi-metal hex phase (cubic → hex at 540 K, hex → liquid at 890 K)	0.5				5×10^{20}	20	10^{-3}

Table 1. Known and estimated values for $Ge_2Sb_2Te_5$,[4] where E_μ, E_n, and E are activation energies of mobility, carrier concentration, and conductivity respectively.

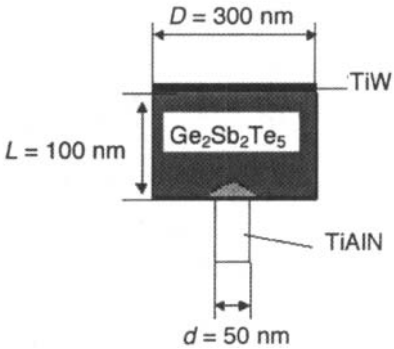


Figure 4. Modern PCM cell of B type.[5]

Latest results on modern PCM cells promise that dimensions down to $D = 10$ nm and $L = 10$ nm may be achievable.[4] If so, this would permit the use of 22 nm design rules, which may be at the limit of Si-based flash memory. The PCM cell design shown in Fig. 4 makes it possible to increase the amorphous part of a cell step by step. Then a multilevel PCM cell may be fabricated. The dependence of the resistance R of such a PCM cell on the set current I_{SET} is shown in Fig. 5.

Let us compare in more detail the performance of new generation PCM cells with Si-based flash memory. The advantages of PCM are the following:

- information retention is ensured by heavy atoms, rather than light electrons;
- simplicity of PCM cell fabrication – a single chalcogenide layer between contacts;
- endurance: Si-based flash memory – 10^5, whereas PCM cells – $10^{11\text{-}12}$;
- set time: Si-based flash requires 10^{-3}–10^{-4} s, whereas PCM cells can be set in ~10^{-7} s (and reset even faster, at ~5×10^{-8} s);

compound	T_{TR} (K)	endurance (cycles)	E_G (eV)
As_2Se_3–$4As_2Te_3$	450	10^4	0.91
As_2Se_3–As_2Te_3	630	10^2	1.17
$3As_2Se_3$–As_2Te_3	1000	1	1.37
As_2Se_3	1500	1	1.90
$Ge_2Sb_2Te_5$		$10^{11\text{-}12}$ Ref. 6	0.74 Ref. 4

Table 2. The composition dependence of semiconductor to metal phase transition temperature T_{TR} and endurance in cycles. The first four rows are from Ref. 7.

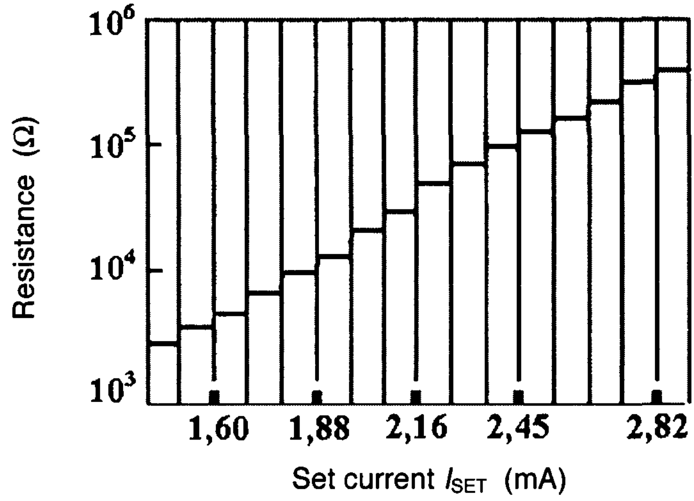


Figure 5. Resistance *vs.* I_{SET} for a multilevel PCM cell.[5]

- scaling: Si-based flash cannot go below the 45 nm design rule due to electron tunneling from the floating electrode, whereas PCM cells down to 10x10 nm may be manufacturable,[4] thus making PCM a possibility for the 22 nm node, unlike Si-based flash;

- PCM or PRAM has higher performance both because the memory element can be switched more quickly and also because single bits may be changed to either 1 (crystalline) or 0 (amorphous) without erasing an entire block of cells.

Conversely, PCM memory suffers from some disadvantages:

- the greatest disadvantage of PCM is the high programming current density ~10^7 A/cm^2, whereas for a typical transistor this value is ~10^5–10^6 A/cm^2;

- another PCM problem is the contact between the hot phase-change region and the adjacent dielectric and metal regions with different thermal expansion properties;

- another challenge for PCM memory is the slow long-term resistance and threshold voltage drift (~$t^{0.1}$);

- density: while PCM cells may eventually reach the 22 nm (or smaller) node, for now the maximum demonstrated integration of PCM is ~0.5 Gb, as opposed to over 8 Gb for flash.

5. Physical model of the set-recess process

In order to further improve PCM memory performance, it is necessary to consider the physical nature of the set-reset processes. Modern PCM cell layout and electrode metal have been chosen so that all cell resistance in the memory state is dominated by the bottom electrode. Then this electrode serves as the "heater" for the reset process. Due to this "heater", a small volume of melted chalcogenide can be quickly cooled to become a glassy semiconductor (short reset pulse of Fig. 3). We shall not consider here the very complex amorphization process, but rather focus on the set process. In this case the cell resistance is due to the high resistance of the CGS material where Joule heating takes place. In this case, it appears likely that the switching and memory effects are determined by electronic-thermal processes. This means that Joule heating, governed by the nonlinear current-voltage characteristic of CGS, is the key to the set process. The nonlinearity takes place in a strong electrical field, and may be described by decreasing of the effective energy of activation $E^* = (E - f(V/V_0))$.

It is known that the electronic-thermal theory of switching can be described by the following equations:

$$Q_1 = Q_2 \quad \text{and} \quad \partial Q_1/\partial T = \partial Q_2/\partial T, \tag{1}$$

where $Q_1 = SL\sigma F^2$, $Q_2 = \lambda S(T - T_0)$, S is the contact area, F is the electric field, λ is the cooling coefficient, and $\sigma = en\mu = \sigma_0\exp(-E^*/kT)$.

For $f(V/V_0) = 0$, we have simple thermal breakdown, $T_{\mathrm{OFF}} = T_{\mathrm{T}} = T_0(1+kT_0/E)$ and $T_{\mathrm{ON}} \sim E/k$, that is the temperature T_{OFF} at the end of the "off" state is very close to the ambient temperature T_0. But the temperature in the "on" state becomes unrealistically high. The situation changes drastically for the electronic-thermal case. For reasons of simplicity, let us consider the case $f(V/V_0) = V/V_0$, where $V_0 = akT$. The experimental results give evidence that two extreme points of the S-shaped current-voltage characteristic (CVC) differ only slightly. In this approximation, the solution of the equations (1) gives the following results: $T_{\mathrm{T}} \sim T_{\mathrm{ON}} \sim 2T_0$, as long as V_{T} and V_{ON} do not differ too much. The exact equality $T_{\mathrm{ON}} = 2T_0$ holds only if the S-shaped CVC disappears entirely when $V_{\mathrm{T}} = V_{\mathrm{ON}}$. A small difference between V_{T} and V_{ON} yields $T_{\mathrm{T}} \leq 2T_0$ and $T_{\mathrm{ON}} \geq 2T_0$, but nevertheless T_{ON} is on the order of $2T_0$ rather than $T_{\mathrm{ON}} \sim E/k$.

Let us consider the case when the phenomenon that determines $f(V/V_0)$ is the multiphonon tunnel ionization of negative-U defects.[8] The existence of such defects in CGS materials is well established. Figure 6(a) shows the model band diagram of the material, whereas Figs. 6(b) and (c) illustrate schematically the multiphonon tunnel ionization of negative-U defects for $D^- - e \rightarrow D^0$ and $D^0 - e \rightarrow D^+$ processes, respectively.

The set equations of the model can be solved,[9] leading to the following approximate result, where τ_2 is characteristic time and q is charge of electron:

$$n = N_{\mathrm{C}} \exp\left(-\left(\frac{E_1 + E_2}{2kT}\right) + \frac{q^2 F^2 \tau_2^3}{3\hbar m} + \frac{\sqrt{2mE_2}}{qF\tau_2} \ln\left[\frac{4q^2 F^2 \tau_2^3}{\hbar m}\right]\right). \tag{2}$$

We find that the last two terms, *i.e.* $f(V/V_0)$, are more complicated than our simple V/V_0 approximation. Still, plugging the carrier concentration (2) in the conductivity $\sigma = en\mu$ and solving equations (1) we obtain a CVC plotted in Fig. 7. We find rather good qualitative agreement between the experimental S-shaped CVC and our theoretical prediction based on the multiphonon tunnel ionization processes of negative-U defects. The 650 °C result obtained for T_{ON} is only ~$3T_0$, where T_0 = 300 K, rather than the much larger E/k.

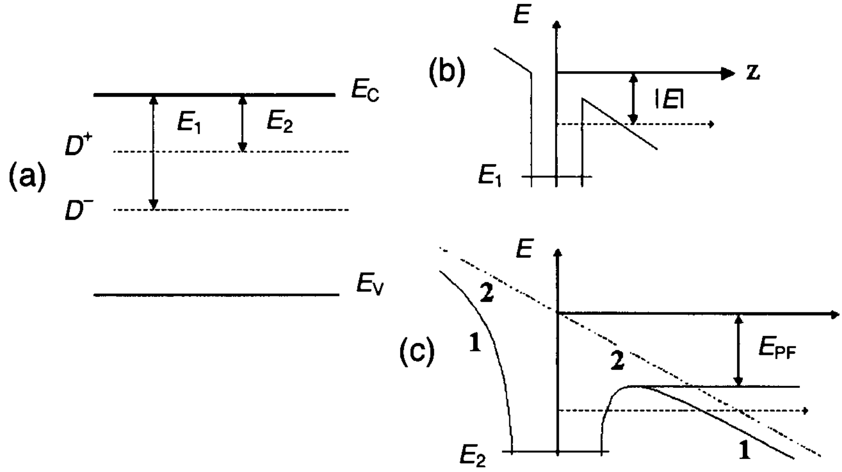


Figure 6. (a) Energy band diagram of the negative-U centers model, E_1 and E_2 are the first and second energies of the negative-U center thermo-ionization; (b) the potential seen by an electron escaping from negative-U center in the D^- state (where $|E|$ is the energy of the escaped electron); (c) the potential seen by an electron escaping from negative-U center in the D^0 state, where line 1 indicates the Coulomb potential, line 2 – the external electric field, and E_{PF} is the Poole-Frenkel energy barrier lowering.

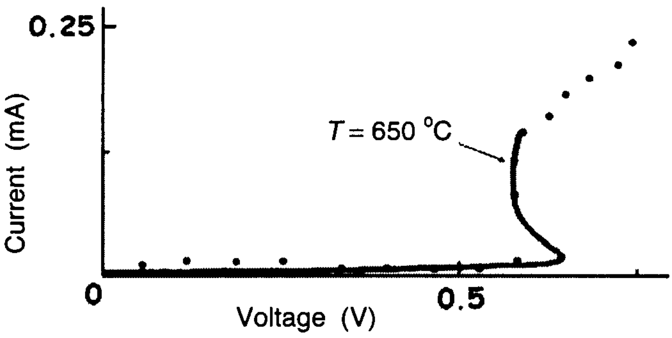


Figure 7. A comparison of experimental[6] (black dots) and theoretical (line) CVCs.

Now, we can formulate from the electronic-thermal theory standpoint the reasons why the new generation PCM cells are successful:

- nanometer dimensions and the rather small difference between V_T and V_{ON} suppresses current filamentation;
- nonlinearity of the current-voltage characteristics decreases effective activation energy of conductivity $E^* = (E - f(V/V_0))$ and hence decreases the on-state temperature T_{ON};[10]
- the initial state is crystalline, so only a small part of the film is made amorphous (reset) and goes back to the crystalline state (set);
- the $Ge_2Sb_2Te_5$ compound provides a large difference between the retention time of the amorphous state at 300 K (tens of years) and the fast, ~100 ns crystallization time at temperature T_C.[6]

All these improvements – nanoscale dimensions of PCM cell, initial crystalline state, and the superior properties of the GST compound – have led to an enormous increase in PCM cell endurance, as shown in Table 2.

5. Conclusions

Today, investment in PCM research has increased sharply. There is an realistic probability that these efforts will succeed in producing manufacturable PCM-based version of flash memory using the CGS compound $Ge_2Sb_2Te_5$.[11,12] Unlike the single element Si, the CGS family consists of an enormous number of compounds. Given a sufficiently accurate theory, one can imagine searching the CGS space for materials with relevant parameters – melting and glassy temperatures, energy activation of crystallization, *etc.* – that will perform even better than $Ge_2Sb_2Te_5$. To summarize, as the title of this chapter notes, it is indeed possible that CGS phase-change memories will replace silicon for nonvolatile applications.

References

1. S. R. Ovshinsky, "Reversible electrical switching phenomena in disordered structures," *Phys. Rev. Lett.* **21**, 1450 (1968).
2. N. A. Gorunova and B. T. Kolomiets, *Zh. Tekhn. Fiz.* **25**, 984 (1955); N. A. Gorunova and B. T. Kolomiets, *Zh. Tekhn. Fiz.* **25**, 2069 (1955).
3. B. T. Kolomiets and E. A. Lebedev, *Radiotekhnika Eletronika* **8**, 2037 (1963).
4. T. Kato and K. Tanaka, "Electronic properties of amorphous and crystalline $Ge_2Sb_2Te_5$ films," *Japan. J. Appl. Phys.* **44**, 7340 (2005).
5. A. I. Popov, *Physics and Technology of Disordered Semiconductors* [in Russian], Moscow: MEI, 2008, p. 282.
6. A. L. Lacaita, "Phase change memories: State of the art, challenges and perspectives," *Solid State Electronics* **50**, 24 (2006).

7. S. A. Smorgonskaya and V. H. Shpunt, Chapter 7 in: K. D. Tsendin, ed., *Electronic Phenomena in Chalcogenide Glassy Semiconductors* [in Russian], St. Petersburg: Nauka, 1996.
8. V. N. Abakumov, V. I. Perel', and I. N. Yassievich, *Nonradiative Recombination in Semiconductors* (*Modern Problems in Condensed Matter Science* series, Vol. 33), Amsterdam: North-Holland, 1991.
9. K. D. Tsendin, "Electro-thermal theory of the switching and memory effects in chalcogenide glassy semiconductors," *phys. stat. sol. (b)* **246**, 1831 (2009)
10. T. Gotoh, "Characteristic at high electric fields in amorphous $Ge_2Sb_2Te_5$ films," *J. Non-Cryst. Solids* **354,** 2728 (2008).
11. A. Popov, "Two generation of phase-change memory devices: difference and common problems," *phys. stat. sol. (b)* **246**, 1837 (2009).
12. A. V. Kolobov, P. Fons, and J. Tominaga, "Local structure of amorphous Ge-Sb-Te alloys: Ge umbrella-flip *vs.* DFT simulations," *phys. stat. sol. (b)* **246,** 1826 (2009).

Current Status and Recent Developments in RSFQ Processor Design

M. Dorojevets
Dept. of Electrical and Computer Engineering
SUNY–Stony Brook, NY 11794-2350, U.S.A.

1. Introduction

The need for high-performance and low power computer systems capable of dealing with the avalanche of data in critical applications is well recognized.

A 2002 report entitled "High Performance Computing for the National Security Community", prepared by NSA in cooperation with other federal agencies, clearly identified the need for user-friendly high-end computing systems with high-bandwidth capabilities because of the existence of critical applications for the national security community "that are neither met nor addressed by the commercial sector".[1]

Some of the key findings of a 2006 report on Joint U.S. Defense Science Board and UK Defense Scientific Advisory Council Task Force on Defense Critical Technologies[2] were that "there are applications that cannot be solved with sufficient speed or with sufficient precision", that "commercial off-the-shelf (COTS) technologies will be insufficient to meet unique military needs", and that the DoD "should invest in critical, defense-niche technologies in order to assure competitive advantage over potential adversaries".

The 2008 DARPA "Exascale Computing Study: Technology Challenges in Achieving Exascale Systems"[3] concluded that there exist four major challenges – energy and power; memory and storage; concurrency and locality; and resiliency, – to achieving exascale systems where current technology trends are simply insufficient, and that significant new research was absolutely needed to bring alternatives on line.

Superconductor processors based on rapid single flux quantum (RSFQ) circuit technology can reach and exceed operating frequencies of 100 GHz, while keeping processor power consumption low. These features provide an opportunity to build different types of systems with ultra-high-speed energy-efficient RSFQ microprocessors to address the government's critical mission needs.

An NSA report entitled "Superconducting Technology Assessment (STA)" was prepared by a group of experts in 2005 and updated in December of 2007.[4] The major conclusion of the report was that the superconducting technology was "sufficiently mature for a major development investment to bring it to a state of readiness for use in high-end machines".

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

2. Past and present RSFQ processor design projects

The issues of RSFQ processor and system design have been addressed in several projects: SPELL processors for the hybrid technology multi-threaded (HTMT) project,[5,6] the 8-bit FLUX processor[7,8] and 8-bit Frontier datapath[9] projects in the U.S., the bit-serial CORE processor[10] and LSRDP computer[11] projects in Japan. These projects are summarized in Table 1.

The HTMT petaflops computer project was a collaboration of several academic, industrial, and U.S. government labs with the goal of studying the feasibility of a petaflops computer system design based on new technologies, including superconductor RSFQ technology.

The HTMT RSFQ-related design work focused on the following issues:

- a multithreaded processor architecture that could tolerate huge disparities between the projected 50–60 GHz speed of RSFQ SPELL processors and the much slower non-superconductor memories outside the cryostat; and
- the projected characteristics of the RSFQ superconductor petaflops subsystem consisting of ~4,000 SPELL processors with a small amount of superconductor memory and a superconductor network for inter-processor communication.

Time Frame	Project	Target Clock	Target peak CPU Performance	Design Status
1997-1999	SPELL processors for the HTMT petaflops system (US)	50-60 GHz	~250 GFLOPS/CPU (est.)	Feasibility study
2000-2002	8-bit FLUX-1 microprocessor prototype (US)	20 GHz	40 billion 8-bit integer operations/s	Designed, fabricated (~63 K JJs, 9.5 mW)
2002-2007	8-bit bit-serial CORE microprocessor prototypes (Japan)	25 GHz bit-serial local, 1 GHz system	~ 250 million 8-bit integer operations/s	Designed, fabricated, and demonstrated (~22K JJs, 6.3 mW)
2007-2011	10 TFLOPS computer with bit-serial large-scale reconfigurable RSFQ datapath (LSRDP)(Japan)	25/60/100 GHz (process dependent)	~ 4 GFLOPS per bit-serial FPU ~2.5 TFLOPS w/ 1024 FPUs (est.)	Work in progress: several bit-serial blocks with up to 11K JJs operating at ~22–25 GHz have been designed and demonstrated
2009-2010	8-bit 20 GHz Frontier processor datapath (US)	20 GHz	20 billion 8-bit operations/second	Work in progress

Table 1. Superconductor RSFQ microprocessor design projects.

The architecture of SPELL processors supports dual-level multithreading with 8–16 multistream units (MSUs), each of which was capable of simultaneous execution of up to four instructions from multiple threads running within each MSU and sharing its set of functional units. However, no processor chip design was done for SPELL processors; their technical characteristics are estimates based on the best projection of RSFQ circuits available at that time (1997–1999).

The 8-bit FLUX-1 microprocessor was the first practical attempt to address architectural and design challenges for 20 GHz RSFQ processors. The FLUX-1 design and fabrication was done in the framework of the FLUX project as a collaboration between the SUNY–Stony Brook, TRW (now Northrop Grumman Space Technology), and NASA's Jet Propulsion Laboratory.

A new partitioned microarchitecture was developed for FLUX-1 with the following distinctive features:

- ultrapipelining to achieve 20 GHz clock rate with only 2–3 Boolean operations per stage;
- two operations per cycle (40 GOPS peak performance for 8-bit data);
- short-distance interaction and reduced connectivity between arithmetic logic units (ALUs) and registers;
- bit-streaming, which allows any operation that is dependent on the result of an operation-in-progress, to start working with the data as soon as its first bit is ready;
- wave pipelining in the instruction memory;
- modular design; and
- ~25 control, integer arithmetic, and logical operations (no load/store operations).

The final FLUX-1 chip, called FLUX-1R, was fabricated in 2002. It had 63,107 Josephson junctions (JJs) on a 10.35x10.65 mm^2 die with power consumption of ~ 9.5 mW at 4.5 K.

Operation of a one-bit ALU-register block (the most complex FLUX-1R component) was confirmed by testing. No operational FLUX-1R chips were demonstrated by the time the project ended in 2002.

Several bit-serial microprocessor prototypes with a very simple architecture called CORE1 have been designed, fabricated, and successfully tested at high speed in the Japanese Superconductor Network Devices project. Participants in this project included Nagoya, Yokohama, and Hokkaido Universities, the National Institute of Information and Communications Technology at Kobe, and the ISTEC Superconductor Research Lab (SRL) at Tsukuba in Japan.

The latest CORE1γ microprocessor has two 8-bit data registers, and two bit-serial ALUs, and small amount (several bytes) of on-chip (shift register type) memory for instructions and data. The instruction set consists of seven 8-bit instructions.

These CORE1 microprocessors have extremely simplified, pipelined processing and control logic, and use slow (1 GHz) system and fast (16–25 GHz) local clocks. The slow 1 GHz system clock is used to advance an instruction from one execution phase to another. The fast local clock is used for bit-serial data transfer and bit-serial data processing within each instruction execution phase. A 25 GHz CORE1γ chip has ~ 22K JJs, and a power consumption of 6.3 mW.

A new long-term large-scale reconfigurable data-path (LSRDP) project with the goal of developing a 10 TFLOPS RSFQ computer started in Japan in 2007. This proposed computer will have a silicon CMOS general-purpose processor and DRAM main memory at room temperature and a cryo-cooled bit-serial RSFQ LRSDP datapath sub-system. The LSRDP section consists of a large number of bit-serial floating-point units connected to each other through a bit-serial programmable operand routing network. No cryo-memory is planned for the LSRDP sub-system. The data flow in LSRDP is one-directional from input to output with no internal information feedback loops.

By 2009, several bit-serial floating point units and routing network chips with up to 11K JJs operating at 22–25 GHz clock frequency have been successfully designed and demonstrated within the framework of the LSRDP project in Japan.

In the meantime, a new RSFQ processor datapath project started in the U.S. in 2009. Its goal is to design and demonstrate the first 8-bit wide 20 GHz processor datapath with on-chip cryo-memory. The datapath will use new asynchronous wave-pipelined wide datapath techniques developed at SUNY–Stony Brook.[12,13] This project is the collaboration between the design teams at Stony Brook and Hypres. The fabrication of the datapath chip using the Hypres 4.5 kA/cm^2 technology is planned for 2010.

3. Major challenges in RSFQ processor design

The availability of ultra-high-speed, low power superconductor circuit technology is only one of several requirements for successful high-performance processor design. Any change in circuit technology calls for a re-examination of the processor microarchitectural and design techniques that work well with it.

The major design challenges are:

- reliable design and synchronization techniques for 50–100 GHz wide datapath RSFQ processors;
- latency tolerance;
- static power reduction;
- memory; and
- CAD tools for VLSI superconductor circuit design.

The processors shown in Table 1 have been designed with use of the canonical RSFQ logic[14] and corresponding cell libraries.[15,16] The canonical RSFQ logic favors super-pipelined designs with very high clock frequencies because almost all logical functions can only be done with clocked gates, each of which logically represents a combination of a stateless Boolean logic element (*e.g.* an inverter) and a flip-flop. As a result, the current synchronous datapath RSFQ designs are optimized for ultra-high processing rates at the expense of significant latency in their long processing pipelines. As is well known, the increase in the clock rate by increasing the number of pipeline stages does not necessarily improve performance.

In order to distribute clock signals to RSFQ cells, binary signal distribution trees built of active elements called pulse splitters have to be used. This leads to unavoidable skew and jitter created in clock distribution networks in synchronous RSFQ processors. The need to synchronize the work of almost all RSFQ gates makes the issue of clock skew and synchronization in general extremely challenging for synchronous RSFQ designs with wide 32/64-bit datapaths. This was one of the reasons (besides the immature technology) why both Flux-1,[7] CORE microprocessors[10] and the LSRDP RSFQ subsystem[11] used bit-serial processing in their ALUs and floating-point units. While this approach allowed the designers to decrease the impact of clock skew and reach frequencies of up to 25 GHz for low-complexity bit-serial processor datapaths, it is not scalable or applicable to future full-fledged 50 GHz 32/64-bit RSFQ processor designs.

The on-chip gate-to-gate communication delays in 50 GHz microprocessors will limit the space reachable in one cycle to ~1–2 mm. The microarchitecture and design of superconductor processors must support a communication-aware, localized computation model in order to have processor functional units fed with data from registers located in close proximity to the units. The most radical attempt to deal with this issue was made in the design of the FLUX-1 RSFQ microprocessor. It had a partitioned microarchitecture where eight ALUs were interleaved with eight registers, which made the distance between them very short. Also, instead of data traveling from registers to ALUs and back, multiple computation waves traveled along static data in registers, which could be called integer bit-streaming processing-in-registers.[7]

Another characteristic of the canonical RSFQ logic is the use of the large bias resistors and static (bias) current to keep junctions biased to ~70% of their critical current value. The static power dissipated in the bias resistors used in each cell is currently an order of magnitude higher than the dynamic power dissipated when junctions in RSFQ cells switch. Although there are known techniques that can significantly reduce static power consumption in RSFQ cells,[17] the issue of static power reduction has not yet been explored in larger complex circuits.

The principal reason for this was the immaturity of the prior fabrication processes capable of fabricating of only relatively low-complexity designs, where the unoptimized static power consumption was still very low. For instance, the most complex 25 GHz CORE1γ processor chip had 22K JJs dissipating 6.3 mW (total power) at cryo-temperatures. Taking into account the cooling costs (from 300 to 5-6 K), the total (wall plug-in) power at room temperature is approximately three orders of magnitude higher than the power required for operation at cryogenic temperatures.

In order to continue to have a competitive edge in energy efficiency over future CMOS designs, RSFQ designers need to switch from the use of canonical RSFQ cell libraries to new libraries designed for low static power consumption. A radical step to completely eliminate bias resistors and static power consumption was proposed recently in a new type of logic called reciprocal quantum logic (RQL).[18]

Cryogenic random access memory (RAM) remains a perennial challenge for superconductor design community. The cryo-memory (more precisely, the lack of it) had the biggest impact on the architecture of superconductor processors and systems shown in Table 1.

In 2007, for the purpose of the new 10 kA/cm^2 fabrication process evaluation, the SRL foundry in Japan fabricated two types of SFQ RAM chips: 4 Kbit and 16 Kbit with 23,488 and 80,768 JJs, respectively. Unfortunately, the yield of 16 Kbit RAM (the largest SFQ RAM implemented by now) was only 63.3% due to the large ac bias current required for memory operation.[19] Besides off-chip RAM, superconductor processors need low-latency, high-throughput pipelined RAM to implement on-chip storage structures capable of working at the same rate as processor functional units.

Finally, the design of a new generation of high-complexity energy-efficient wide-datapath RSFQ processors with clock frequencies reaching 50 GHz creates additional requirements for CAD tools. For the cell-level design, this requires the development of new VHDL/Verilog cell libraries capable of taking into account gate delay fluctuations, and measuring both static and dynamic power consumption. For instance, the reduction of static current through the use of the LR-load biasing technique[17] will make the cell propagation delays slightly dependent on the operation of neighboring cells sharing the same bias lines. This may require the context-dependent calculation of propagation delays for each cell.

4. Asynchronous hybrid wave pipelining for 50 GHz 32/64-bit RSFQ processors

Efficient synchronization and design techniques are among the biggest challenges for the 50 GHz wide datapath RSFQ processor design.

In conventional synchronous pipelining, intermediate pipeline latches are used in addition to input and output registers. Intermediate latches ensure that data get propagated from one stage to another in a synchronous manner. There is only one set of data between pipeline stages. The clock cycle time in conventional pipelining is determined by the worst-case operation of the stage with the largest delay plus pipeline synchronization overhead, which can be defined as a sum of the latch propagation delay, latch set-up time, and clock skew between adjacent stages. In conventional synchronous RSFQ pipelines, this synchronization overhead can easily exceed latency from the logic blocks due to the small amount of work per stage, and significant clock skew. Making the situation even more difficult is the unavoidable (temperature-induced) timing uncertainty in RSFQ gate delays.

Wave pipelining is an approach aimed to achieve high performance in pipelined systems by removing intermediate latches.[20,21] This technique increases the clock frequency by allowing multiple data waves to exist in any stage. By removing intermediate registers, the area, power and load associated with the clock are reduced. The clock cycle time in wave pipelined systems is determined by the difference between the maximum and minimum delays through the combinational logic (plus register set-up and hold times plus double clock skew). The challenges of designing wave pipelined systems are twofold:

- preventing collision of unrelated data waves; and
- balancing (equalizing) delay paths in order to reduce differences between the longest and shortest delays through the combinational logic.

The differences can accumulate as the waves propagate through a pipeline, creating the potential for data overrun of unrelated data waves. Ultra-high-rate long RSFQ pipelines are especially prone to this problem.

These problems can be avoided by using the hybrid wave-pipelining approach, where signals are held so that the next stage does not start operating until all the signals from the previous stage are available. In the hybrid wave-pipelined datapath, the stage with the largest delay difference determines the clock cycle time of the datapath.

Many RSFQ designs (*e.g.* the FLUX-1 processor pipeline) used some form of hybrid wave pipelining known as co-flow synchronization, with the clock traveling with data across the pipeline, thus eliminating the need for complex central clock distribution. The challenge for this co-flow synchronization is the necessity to insert carefully calculated delays into the clock propagation path to honor set-up and hold time requirements for clocked RSFQ gates. The problem becomes more and more severe as the width, length, and complexity of the datapath grow, and timing uncertainty increases.

A new asynchronous hybrid wave-pipelined (AHWP) methodology has been developed at Stony Brook University[9] to remove the necessity of centralized clocking as in classic synchronous pipelining and the need to propagate clock signals along the data as in traditional wave pipelining. In the AHWP pipelines, the moment when pipeline stage logic starts processing will be determined by the availability of data, not clock signals. Thus, AHWP pipelines are truly data-driven.

This approach has an evident advantage of eliminating the problem of clock distribution within processor pipeline. The problem of how to detect the event when all data are available becomes the key design issue for control logic at each stage. This problem can be solved by appropriate coding of data, *e.g.* dual-rail, and the use of reset signals that "clean" the logic of each pipeline stage immediately after the propagation of data waves through the stage.

Another typical design problem that needs to be addressed is how and where the data that arrive earlier can wait for other signals to catch up. In current clocked RSFQ gates, input data signals (pulses) can "wait" inside the gates until the clock signal arrives and produces an output pulse, if any. Unfortunately, these gates need a clock signal to produce an output, so they alone are not suited for data-driven (push-through by data) asynchronous computing. It should be mentioned that there are other asynchronous techniques and gates for RSFQ such as delay-insensitive RSFQ logic, which is based on dual-rail signal coding.[22] A drawback of delay-insensitive logic is that it requires a lot of junctions for its implementation and is basically not resettable (at least in its current form.)

The AHWP design methodology includes the use of a tunable generic VHDL gate library and statistical modeling of the gate delay variations. The technology-tunable parameters of the VHDL gate models will include mean delays, RMS jitter, power consumption, as well as the complexity of each gate in terms of JTLs. The VHDL gate models are designed to simulate normally-distributed thermally-induced gate delay variations for a given value of delay variance.

The preliminary study of the efficiency of hybrid wave-pipelining for implementation of several arithmetic and data routing blocks have been carried out.[9,10] The results show that hybrid wave-pipelining can tolerate the four-fold increase in RSFQ datapath width from 16 to 64 bits with insignificant (~4–5%) decrease in processing rate. This technique can be successfully used in regular circuits where balancing of delay paths for data can be done relatively easily.

It should be noted, however, that a full-fledged 50 GHz RSFQ processor would use a combination of several asynchronous and synchronous techniques. Among the key reasons for this are the irregularity of control (*e.g.* instruction issue) logic and the need to control the injection rate at which data and operations can be safely "pumped" into wave-pipelined processor units.

All of this means that the 32/64-bit RSFQ processor microarchitecture needs to be developed in a way that allows both efficient and reliable coexistence of synchronous and asynchronous components in the processor datapath.

5. Conclusions

During the last several years, designers were able to identify and demonstrate solutions to several design and fabrication problems related to bias current shielding, 100 GHz passive transmission line design, 25 GHz operational frequencies with relatively wide operational margins for designs with more than 22K JJs, *etc.* A new 1.0 μm 10 kA/cm^2 8 Nb metal layer technology developed in Japan allowed the fabrication of chips with the complexity of up to 10^5 JJs. Recent design and architectural studies have demonstrated that new synchronization techniques, such as asynchronous wave-pipelining, can be very efficient in dealing with jitter and gate delay fluctuations in future wide datapath RSFQ processors.

These recent developments in fabrication, design, and microarchitecture created an opportunity for an important step towards full-fledged superconductor RSFQ processors. RSFQ designers need to exploit this opportunity and demonstrate sufficient functionality, complexity, reliability, speed, and energy efficiency in practical designs in order for superconductor processors to be seriously considered real competitors to CMOS.

Acknowledgments

This work was supported by the US Department of Defense under an ARO contract.

References

1. "High performance computing for the national security community," NSA report (July, 2002), see *http://www.nitrd.gov/pubs/nsa/sta.pdf*
2. Joint U.S. Defense Science Board and UK Defense Scientific Advisory Council Task Force on Defense Critical Technologies report (2006), see *http://www.acq.osd.mil/dsb/reports/2006-03-Defense_Critical_Technologies.pdf*
3. "Exascale computing study: Technology challenges in achieving exascale systems," DARPA report (2008), see *http://users.ece.gatech.edu/~mrichard/ExascaleComputingStudyReports/exascale_final_report_100208.pdf*
4. "Superconducting technology assessment," NSA NITRD report (August, 2005); "Superconducting technology assessment (update)," NSA NITRD report (December, 2007).
5. M. Dorojevets, P. Bunyk, D. Zinoviev, and K. Likharev, "COOL-0: design of an RSFQ subsystem for petaflops computing," *IEEE Trans. Appl. Supercond.* **9**, 3606 (1999).
6. M. Dorojevets, "COOL multithreading in HTMT SPELL-1 processors," chapter in: Y. S. Park, S. Luryi, M. S. Shur, J. M. Xu, and A. Zaslavsky, eds., *Frontiers in Electronics: From Materials to Systems*, Singapore: World Scientific Publishing, 2000, pp. 247–253.
7. M. Dorojevets, "An 8-bit FLUX-1 RSFQ microprocessor built in 1.75-μm technology," *Physica C* **378-381**, 1446 (2002).
8. M. Dorojevets, D. Strukov, A. Silver, *et al.*, "On the road towards superconductor computers: Twenty years later," chapter in S. Luryi, J. M. Xu, and A. Zaslavsky, eds., *Future Trends in Microelectronics: The Nano, the Giga, and the Ultra*, New York: Wiley, 2004, pp. 214–225.
9. US ARO RDECON ACQ CTR contract W911NF-0-C-0036 awarded to Hypres, 2009.
10. A. Fujimaki, M. Tanaka, T. Yamada, Y. Yamanashi, H. Park, and N. Yoshikawa. "Bit-serial single flux quantum microprocessor CORE," *IEICE Trans. Electronics* **E91-C**, 342 (2008).
11. N. Takagi, K. Murakami, A. Fujimaki, N. Yoshikawa, and K Inoue, "Proposal of a desk-side supercomputer with reconfigurable data-paths using rapid single-flux-quantum circuits," *IEICE Trans. Electronics* **E91-C**, 350 (2008).

12. M. Dorojevets, C. Ayala, and A. Kasperek, "Development and evaluation of design techniques for high-performance wave-pipelined wide datapath RSFQ processors," *Proc. Intern. Supercond. Electronics Conf.* (2009), SP-P46.
13. M. Dorojevets and C. Ayala, "Logical design and analysis of a 32/64-bit wave-pipelined RSFQ adder," *Proc. Supercond. SFQ VLSI Workshop* (2009), pp. 15–16.
14. K. Likharev and V. Semenov, "RSFQ logic/memory family: A new Josephson junction technology for sub-terahertz clock frequency digital systems," *IEEE Trans. Appl. Supercond.* **1**, 3 (1991).
15 P. Bunyk, K. Likharev, and D. Zinoviev, "RSFQ technology: Physics and devices," *Intern. J. High Speed Electronic Syst.* **11**, 257 (2001).
16. S. Yorozu, Y. Kameda, H. Terai, A. Fujimaki, T. Yamada, and S. Tahara, "A single flux quantum standard logic cell library," *Physica C* **378-381**, 1471 (2002).
17. T. Nishigai, S. Yamada, and N. Yoshikawa, "Design and implementation of low-power SFQ circuits using LR-load biasing technique," *Physica C* **445-448**, 1029 (2006).
18. Q. P. Herr and A. Y. Herr, "Reciprocal quantum logic," *Proc. ASC* (2008), Chicago, U.S.A.
19. M. Hidaka, S. Nagasawa, K. Hinode, and T. Satoh, "Improvements in fabrication process for Nb-based single flux quantum circuits in Japan," *IEICE Trans. Electronics* **E91-C**, 318 (2008).
20. L. W. Cotton, "Maximum-rate pipeline systems," in *Proc. Spring Joint Computer Conf.* (1969), pp. 581-586.
21. W. P. Burleson, M. Ciesielski, F. Klass, and W. Liu, "Wave pipelining: A tutorial and research survey," *IEEE Trans. VLSI Syst.* **6**, 464 (1998).
22. P. Patra, S. Polonsky, and D. Fussel, "Delay insensitive logic for RSFQ superconductor computing," *Proc. 3rd Intern. Symp. Advanced Res. Async. Circ. Syst.* (1997), pp. 42–51.

1/*f* Noise: The Funeral is Cancelled (or Postponed)

M. E. Levinshtein and S. L. Rumyantsev[*]
Ioffe Physical-Technical Institute of Russian Academy of Sciences
St. Petersburg, Russia

1. Introduction

Some years ago, a talk proposing the end of the 1/*f* noise era was presented at the "Future Trends in Microelectronics: the Nano Millennium" workshop on Ile de Bendor (2001). This talk was based on numerous papers where disappearance of the 1/*f* noise was predicted in devices "of the next generation". It was stated that 1/*f* noise would vanish as the size of "future" semiconductor devices decreased.

It looked like this prediction was based on a very firm foundation. Indeed, it is widely believed that 1/*f* noise in semiconductors (and metals) appears as a superposition of random processes with very wide distribution of characteristic times τ_i. However, if the device is extremely small, there can be only one or two active defects causing these random processes. In this case, the noise in the time domain is of the random telegraph signal (RTS) form. The spectrum of this noise is single Lorentzian (one defect) or superposition of two Lorentzians (two defects).

Figure 1 shows an example of the noise spectra for three small MOSFETs[1] with the gate area $A = 0.4\ \mu m^2$. Since downscaling the device size is the mainstay of semiconductor technology, it was expected that the 1/*f* noise would vanish in future, smaller semiconductor devices. The "future" is now the present and, surprisingly, the prediction has failed.

2. Low frequency noise in modern semiconductor devices.

Figure 2 shows cross-sectional view of a modern InAlAs/InGaAs MODFET with gate length L_G = 60 nm and gate width W = 20 μm. The InAlAs/InGaAs heterostructure was grown by molecular beam epitaxy (MBE) on a semi-insulating InP substrate. The structure consists of a 200-nm-thick $In_{0.52}Al_{0.48}As$ buffer layer; 15 nm thick undoped $In_{0.7}Ga_{0.3}As$ pseudomorphic channel layer; followed by 5 nm thick $In_{0.52}Al_{0.48}As$ spacer; δ-doped layer (Si dopant, 6×10^{12} cm^{-2}); 12 nm thick $In_{0.52}Al_{0.48}As$ barrier layer; and a 10 nm thick $In_{53}Ga_{47}As$ cap layer Si-doped to 6×10^{18} cm^{-3}.[2]

Hall measurements yielded the electron concentration in regions protected by the cap layer to be ~5.5×10^{12} cm^{-2}, whereas in parts of the channel not covered by

---

[*] Also with Rensselaer Polytechnic Institute, Troy, NY 12180-3590, U.S.A.

Copyright © 2010 John Wiley & Sons, Inc.

the cap layer (see Fig. 2) the electron density was ~3.5×10^{12} cm^{-2}. Electron Hall mobility in large samples was within the 11000–12200 cm^2/V·s range.

The low-frequency noise was measured on the packaged transistors in the 1 Hz to 50 kHz frequency range at 300 K with the source grounded.

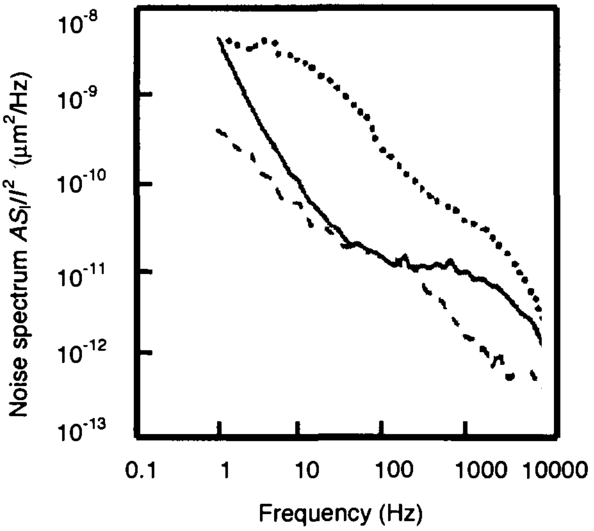


Figure 1. Noise spectra of three MOSFETs with active area of 0.4 μm^2. Total number of the carriers in these samples is $N \sim 10^3$ (from Ref. 1).

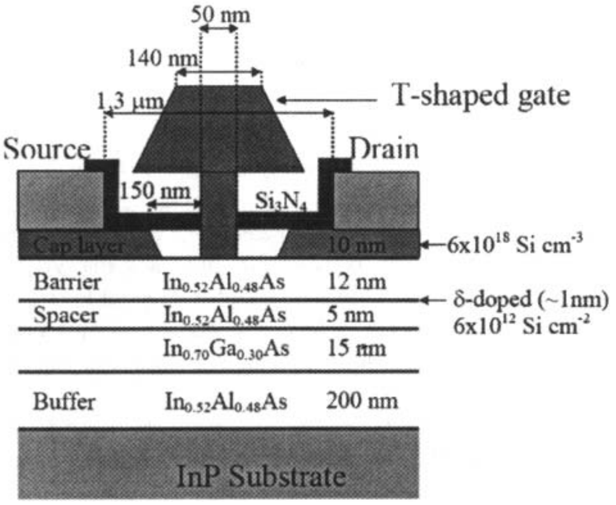


Figure 2. Cross-sectional schematic view of an InAlAs/InGaAs MODFET with L_G = 60 nm, W = 20 μm (from Ref. 2).

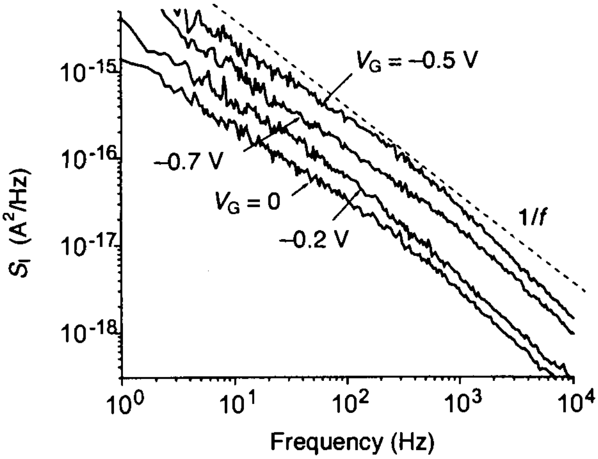


Figure 3. Frequency dependences of drain current fluctuations S_I in linear regime at different gate voltages V_G. Dashed line shows the 1/*f* slope.[2]

Figure 3 demonstrates noise spectra of the drain current fluctuations at different gate voltage V_G. At $V_G = 0$, the number of carriers in the channel $N \sim 31000$, whereas at $V_G = -0.7$ V, $N \sim 3800$. As can be seen from Fig. 3, the noise spectra demonstrate obvious 1/*f* behavior even at the smallest N value.

A recent publication reported the noise in GaN nanowires.[3] Figure 4 shows field emission scanning electron microscopy (FESEM) images of GaN nanowires in the growth matrix, with diameter ranging from 50 to 200 nm. These GaN nanowires were processed into field effect transistors with a common gate (Si substrate) and symmetrically designed drain and source contacts (see inset in Fig. 5).

The number of carriers in the channel depended on substrate bias V_G, dropping to $N \sim 4\mathrm{x}10^3$ at $V_G = 5.5$ V. As seen in Fig. 5, noise spectra have 1/*f* form (with some signature of separate Lorentzians in some samples).

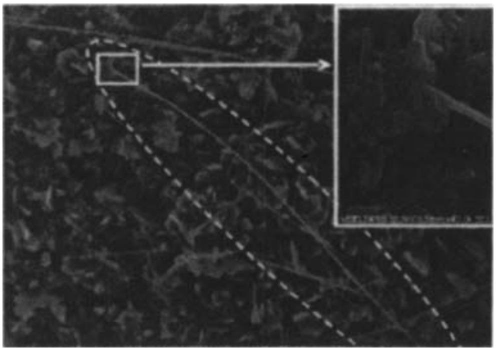


Figure 4. Field emission SEM images of GaN nanowires in the growth matrix. The inset shows a higher magnification image of the GaN nanowires.[3]

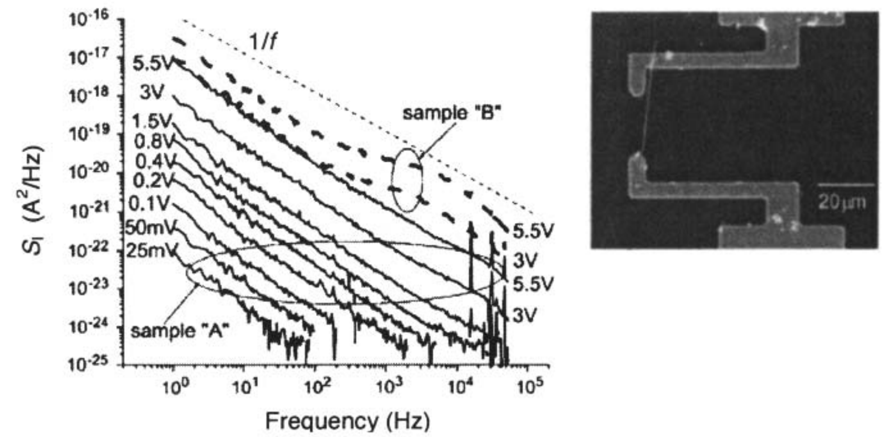


Figure 5. Frequency dependence of drain current fluctuations S_I in the linear regime for different V_G values in a GaN nanowire transistor.[3]

The low frequency noise in single-walled carbon nanotubes (SWNTs), both metallic (M-SWNTs) and semiconducting (S-SWNTs), deposited on a Si substrate has also been reported – see Fig. 6.[4] Again, the frequency dependence of spectral noise density is of the 1/*f* form. No decomposition into separate Lorentzians has been observed even at substrate bias at which current through S-SWNTs was smaller than 10^{-11} A.

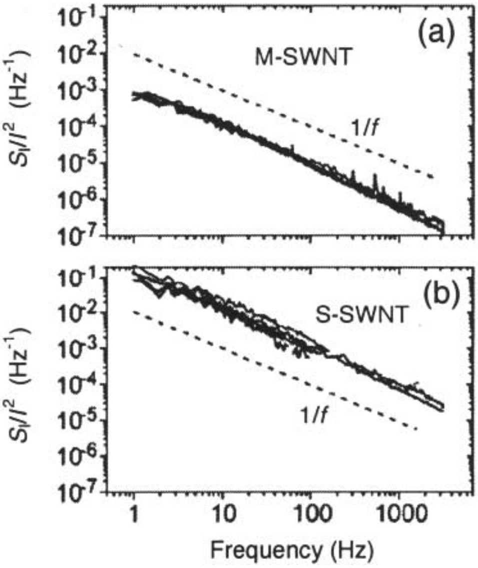


Figure 6. Frequency dependence of the relative current noise S_I/I^2 in the linear regime for M-SWNTs (a) and S-SWNTs (b).[4]

Low frequency noise was investigated in graphene single layer devices of 0.5 nm thickness, 1 μm length, and 20–200 nm width.[5] The active volume V of smallest sample was about 10^{-17} cm^{-3}.

Normalized current noise spectra S_I/I^2 for two graphene devices with different widths, 20 and 200 nm, are shown in Fig. 7. The noise spectra show the characteristic 1/*f* dependence. In the 200 nm device, the noise amplitude is ~10 times lower than in the 20 nm device, in agreement with $1/V$ behavior, indicating that edge state scattering does not dominate the 1/*f* noise.

Current noise power S_I at $f = 10$ Hz as a function of current I is shown in Figure 8 for the 20 nm wide graphene single layer device for different drain voltages V_D. The observed I^2 dependence indicates that current fluctuations are a manifestation of "volume" resistance fluctuations.

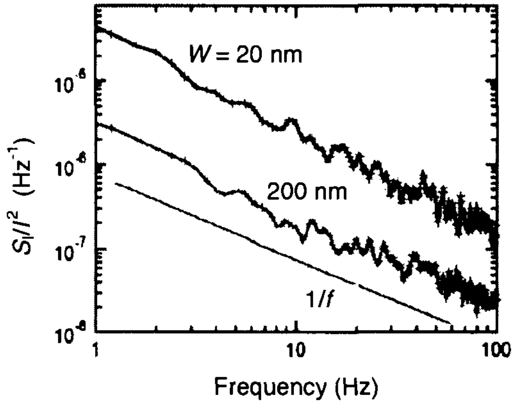


Figure 7. Frequency dependences of relative current noise S_I/I^2 for single-layer graphene nanoribbons of 20 and 200 nm width.

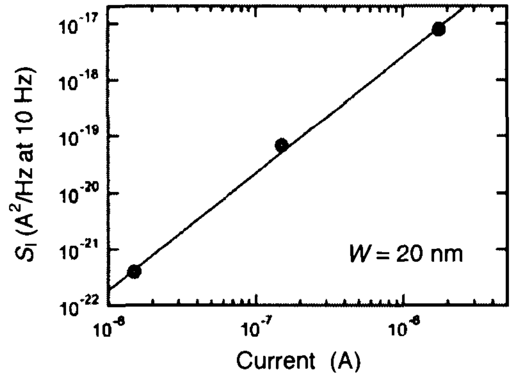


Figure 8. Current noise power S_I at $f = 10$ Hz *vs.* current I for a 20 nm wide single-layer graphene device. Drain voltages V_D = 10 mV, 100 mV and 1 V. The current noise $S_I \sim I^2$, indicating it is caused by resistance fluctuations.

Modern MOSFETs also demonstrate 1/*f* noise behavior. As mentioned, for example, in Ref. 6, the drain current noise of MOSFETs with L_G = 30 nm and W = 10 μm (area A = 0.3 μm^2) "shows a clear 1/*f* noise behavior over which some Lorentzian G–R components may superimpose".

As seen from results presented here, 1/*f* noise is observed in extremely small modern devices and therefore the funeral of 1/*f* noise is cancelled (or postponed).

3. Analysis and discussion

As seen in Fig. 1, if the gate area decreases and all other parameters are kept constant, the 1/*f* noise decomposes into separate Lorentzians. However, in any real device *scaling rules* dictate that other quantities should also change.

In MOSFETs, for example, a decrease in gate length L_G must be accompanied by increase in channel electron concentration n (typically $n \sim 1/L_G$, see, for example, Ref. 7). In the meantime, there is a well-established empirical rule that the concentration of defects n_T is proportional (more or less) to the doping concentration n_0. (That is not a decree of nature, but rather decree of technology). For this reason, the number of the defects in the channel remains virtually invariable as L_G decreases: $n_T \sim n_0 \sim 1/L_G$, keeping the $n_T L_G$ product constant.

Besides, according to scaling rules, decrease in gate length L_G must be accompanied by a decrease in dielectric thickness t_{INS}. In modern MOSFETs with $L_G \leq 30$ nm, t_{INS} is down to a few nanometers (just several tens of atomic layers). Leakage current through such thin layers increases sharply and contributes to the total noise. Concentration of defects in such thin layers is much larger than that in "thick" oxide layers. Attempts to use thicker layers with higher dielectric constant (high-κ layers) make the noise situation even worse: high-κ layers have higher trap concentrations which causes higher 1/*f* noise.

Furthermore, it is necessary to take into account that as the size of the active region decreases, it becomes ever more difficult to discriminate between the "volume" and the "surface" of the device. Even if a very thin (several nanometers) "bulk" layer is totally defect-free, the surface states cause low frequency noise due to Coulomb interaction, surface roughness, and tunneling from "volume" to "surface" states.

Therefore, the funeral of the 1/*f* noise is at least postponed.

4. Conclusion

The prediction made several years ago about the disappearance of 1/*f* noise as "future" semiconductor devices decrease in size has failed. The main reason for this is the increase in defect concentrations due to downscaling of the devices. The use of new materials and processes also increases the defect concentration.

However, throughout the history of semiconductor devices, technological advances have gradually decreased the volume and surface defect densities. For example, Si technology started with 10^{15} cm^{-2} surface trap density decades ago, but nowadays can boast densities down to ~10^{9} cm^{-2}. Therefore, we can hope that the trap density in highly doped layers and high-κ dielectrics will decrease in the future. That gives us a hope for the end of the 1/*f* era in semiconductor devices.

Acknowledgments

This work was supported by Russian Foundation for Basic Research (grant 08-02-00010)

References

1. M. J. Uren, D. J. Day, and M. J. Kirton, "1/*f* and random telegraph noise in silicon metal-oxide-semiconductor field-effect transistors," *Appl. Phys. Lett.* **47**, 1196 (1985).
2. M. E. Levinshtein, S. L. Rumyantsev, R. Tauk, *et al.*, "Low frequency noise in InAlAs/InGaAs modulation doped field effect transistors with 50-nm gate length," *J. Appl. Phys.* **102**, 064506 (2007).
3. S. L. Rumyantsev, M. S. Shur, M. E. Levinshtein, A. Motayed, and A. V. Davydov, "Low-frequency noise in GaN nanowire transistors," *J. Appl. Phys.* **103**, 064501 (2008).
4. A. Vijayaraghavan, S. Kar, S. Rumyantsev, *et al.*, "Effect of ambient pressure on resistance and resistance fluctuations in single-wall carbon nanotube devices," *J. Appl. Phys.* **100**, 024315 (2006).
5. Z. Chen, Y. M. Lin, M. J. Rooks, and Ph. Avouris, "Graphene nano-ribbon electronics," *Physica E* **40**, 228 (2007).
6. T. Contaret, K. Romanjek, T. Boutchacha, G. Ghibaudo, and F. Boeuf, "Low frequency noise characterization and modeling in ultrathin oxide MOSFETs," *Solid State Electronics* **50**, 63 (2006).
7. M. S. Shur, *Physics of Semiconductor Devices*, New Jersey: Prentice Hall, 1990, pp. 353–396.

This page intentionally left blank

Part III

Physics, Biology, and Other Sister Sciences

This page intentionally left blank

3 Physics, Biology, and Other Sister Sciences

While fashionable trends like quantum computing and spintronics have always been subjected to critical examinations at the FTM meetings, they also continued to enjoy insightful analyses of effects that can be created or harvested based on fundamental physics.

While technological limits of proposals of using spin and quantum bits for computing were further revealed, fundamental phenomena such as the spin Hall effect, and long spin coherence length in silicon, and biological processes of memory allocation and recall at the cellular level received renewed attention.

In this part of the volume, participants also address new systems such as graphene and self-assembled nanostructures.

Contributors

3.1	M. I. Dyakonov
3.2	W. E. van den Berg and S. A. Kushner
3.3	V. Shelukhin, M. Karpovski, A. Palevski, J. Xia, A. Kapitulnik, and A. Tsukernik
3.4	V. Sverdlov, O. Baumgartner, T. Windbacher, and S. Selberherr
3.5	V. Ryzhii, M. Ryzhii, A. Satou, N. Ryabova, T. Otsuji, V. Mitin, F. T. Vasko, A. A. Dubinov, V. Y. Aleshkin, and M. S. Shur
3.6	R. L. Williams, D. Dalacu, M. E. Reimer, K. Mnaymneh, V. Sazonova, P. J. Poole, G. C. Aers, R. Cheriton, S. Frédérick, D. Kim, J. Lapointe, P. Hawrylak, and M. Korkusiński

Spin Hall Effect

M. I. Dyakonov
Université Montpellier II, CNRS, 34095 Montpellier, France

1. Introduction

The spin Hall effect (SHE) originates from the coupling of the charge and spin currents due to spin-orbit interaction. It was predicted in 1971 by Dyakonov and Perel'.[1,2] Following the suggestion in Ref. 3, the first experiments in this domain were done by Fleisher's group at the Ioffe Institute in St. Petersburg,[4,5] providing the first observation of what is now called the inverse SHE. As for the SHE itself, it had to wait for 33 years before it was experimentally observed by two groups in Santa Barbara[6] and in Cambridge, UK.[7] These observations aroused considerable interest and triggered intense research, both experimental and theoretical, with hundreds of publications.

The spin Hall effect consists in spin accumulation at the lateral boundaries of a current-carrying conductor, the directions of the spins being opposite at the opposing boundaries, see Fig. 1. For a cylindrical wire the spins wind around the surface. The boundary spin polarization is proportional to the current and changes sign when the direction of the current is reversed.

The term "spin Hall effect" was introduced by Hirsch in 1998.[8] It is indeed somewhat similar to the normal Hall effect, where *charges* of opposite signs accumulate at the sample boundaries due to the action of the Lorentz force in a magnetic field. However, there are significant differences. First, no magnetic

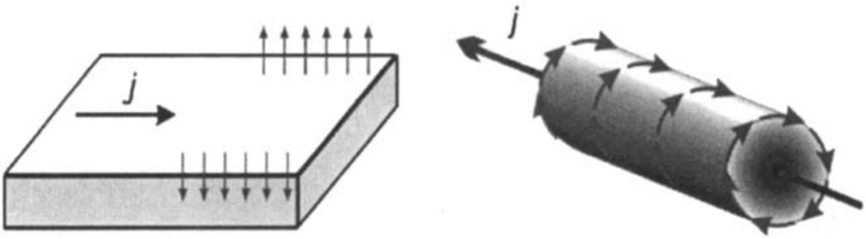


Figure 1. The spin Hall effect. An electric current induces spin accumulation at the lateral boundaries of the sample. In a cylindrical wire the spins wind around the surface, like the lines of the magnetic field produced by the current. However the value of the spin polarization is much greater than the (usually negligible) equilibrium spin polarization in this magnetic field.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

field is needed for spin accumulation. On the contrary, if a magnetic field perpendicular to the spin direction is applied, it will destroy the spin polarization. Second, the value of the spin polarization at the boundaries is limited by spin relaxation, and the polarization exists in relatively wide *spin layers* determined by the spin diffusion length, typically on the order of 1 μm (as opposed to the much smaller Debye screening length that determines the charge accumulation in the normal Hall effect).

In this chapter, we will discuss the phenomenology of spin-charge current coupling, which is true whatever is the microscopic mechanism involved. We will also consider qualitatively the physical origin of various effects resulting from spin asymmetry in electron scattering, in particular the previously unknown phenomenon of *swapping* of spin currents. A more detailed discussion can be found in Ref. 9.

2. Phenomenology

We will now discuss, from pure symmetry considerations, what phenomena of spin-charge coupling are possible in principle. Here we restrict ourselves to isotropic media with inversion symmetry. This does not mean that the results presented below are not valid when inversion symmetry is absent. Rather, we will not take into account additional specific effects that are due entirely to the lack of inversion symmetry. Our phenomenological approach allows us to describe a number of interesting physical effects by introducing only two dimensionless parameters.

- *Spin currents*

The notion of spin current was introduced in Ref. 1. It is described by a tensor q_{ij}, where the first index indicates the direction of flow, while the second denotes which component of the spin is flowing. Thus, if all electrons with concentration n are completely spin-polarized along z and move with a velocity v in the x direction, the only non-zero component of q_{ij} is $q_{xz} = nv$. (Since $s = 1/2$, it might be more natural to define the spin current density for this case as $nv/2$. However, it is more convenient to omit 1/2, because this allows avoiding numerous factors 1/2 and 2 in other places. It would be more correct to describe our definition of q_{ij} as the spin *polarization* current density tensor. Below we will use the shorthand "spin current".)

Both the charge current and the spin current change sign under space inversion (because spin is a pseudo-vector). By contrast, they behave differently with respect to time inversion: while the electric current changes sign, the spin current does not (because spin, like velocity, changes sign under time inversion).

- *Coupling of spin and charge currents*

The transport phenomena related to coupling of the spin and charge currents can be described phenomenologically in the following simple way.[9,10] We introduce the charge and spin flow densities, $\boldsymbol{q}^{(0)}$ and $q_{ij}^{(0)}$, which would exist in the absence of spin-orbit interaction:

$$q_i^{(0)} = -\mu n E_i - D\partial n/\partial x_i\,, \tag{1}$$

$$q_{ij}^{(0)} = -\mu n E_i P_j - D\partial P_j/\partial x_i\,, \tag{2}$$

where μ and D are the mobility and diffusivity, connected by the Einstein relation, n is the electron concentration, $\boldsymbol{E}$ is the electric field, and $\boldsymbol{P}$ is the spin polarization density (it is convenient to use this quantity, instead of the normal spin density $\boldsymbol{S} = \boldsymbol{P}/2$).

Equation (1) is the standard drift-diffusion expression for the electron flow, while Eq. (2) describes the spin current of polarized electrons, which may exist even in the absence of spin-orbit interaction, simply because spins are carried by the electron flow. We ignore the possible dependence of mobility on spin polarization, which is assumed to be small. If there are other sources for currents, for example a temperature gradient, the corresponding terms should be included in Eqs. (1) and (2).

Spin-orbit interaction couples the two currents and gives corrections to the values of the primary currents $\boldsymbol{q}^{(0)}$ and $q_{ij}^{(0)}$. For an isotropic material with inversion symmetry, we have:[10]

$$q_i = q_i^{(0)} + \gamma\varepsilon_{ijk} q_{jk}^{(0)}\,, \tag{3}$$

$$q_{ij} = q_{ij}^{(0)} - \gamma\varepsilon_{ijk} q_k^{(0)}\,, \tag{4}$$

where q_i and q_{ij} are the corrected currents, ε_{ijk} is the unit antisymmetric tensor (whose non-zero components are $\varepsilon_{xyz} = \varepsilon_{zxy} = \varepsilon_{yxz} = -\varepsilon_{yxz} = -\varepsilon_{zyx} = -\varepsilon_{xzy} = 1$), and γ is the small dimensionless parameter proportional to the strength of the spin-orbit interaction. Sums over repeating indices are assumed. The difference in the signs in Eqs. (3) and (4) is due to the different properties of charge and spin currents with respect to time inversion.

Thus, a q_{xy} spin current induces a charge flow in the z direction (q_z) and, conversely, a flow of charge in the z direction induces spin currents q_{xy} and q_{yx}. This transformation is analogous to the Magnus effect: the deviation of a spinning tennis ball from its straight path in air depends on the sense of rotation. The same is true for electrons, but for another reason, the spin-orbit interaction.

In fact, as noted already in Refs. 1 and 2, symmetry considerations allow for some additional terms in Eq. (4), to be discussed below.

3. Phenomenological equations

Explicit phenomenological expressions for the two currents follow from Eqs. (1)-(4) (the electric current density $\boldsymbol{j}$ is related to $\boldsymbol{q}$ by $\boldsymbol{j} = -e\boldsymbol{q}$, where e is the elementary charge):

$$\boldsymbol{j}/e = \mu n\boldsymbol{E} + D\nabla n + \beta \boldsymbol{E} \mathrm{x} \boldsymbol{P} + \delta \nabla \mathrm{x} \boldsymbol{P}\,, \tag{5}$$

$$q_{\mathrm{ij}} = -\mu n E_{\mathrm{i}} P_{\mathrm{j}} - D\partial P_{\mathrm{j}}/\partial x_{\mathrm{i}} + \varepsilon_{\mathrm{ijk}}(\beta n E_{\mathrm{k}} + \delta \partial n/\partial x_{\mathrm{k}}) \tag{6}$$

Here $\beta = \gamma\mu$ and $\delta = \gamma D$, so that the coefficients β and δ, similar to μ and D, satisfy the Einstein relation. However, since γ changes sign under time inversion, β and δ are *non-dissipative* kinetic coefficients, unlike μ and D.

Equations (5) and (6) should be complemented by the continuity equation for the spin polarization vector:

$$\partial P_{\mathrm{j}}/\partial t = \partial q_{\mathrm{ij}}/\partial x_{\mathrm{i}} + P_{\mathrm{j}}/\tau_{\mathrm{S}}\,, \tag{7}$$

where τ_{S} is the spin relaxation time.

While Eqs. (5)-(7) are written for three dimensions, they are equally applicable to the 2D case, with obvious modifications: the electric field, space gradients, and all currents (but not the spin polarization vector) should have components in the 2D plane only.

4. Physical consequences of spin-charge coupling

Equations (5)–(7), which appeared for the first time in Refs. 1 and 2, describe the physical consequences of spin-charge current coupling. The effects of spin-orbit interaction are contained in the additional terms with the coefficients β and δ. Let us now consider these additional terms.

- *Anomalous Hall effect*

The term $\beta \boldsymbol{E}\mathrm{x}\boldsymbol{P}$ in Eq. (5) describes the anomalous Hall effect, which is observed in ferromagnets and has been known for a very long time (it was first observed by Hall himself in 1881). The measured Hall voltage contains a part that is proportional to magnetization, but cannot be explained as being due to the magnetic field produced by magnetization (it is much greater than that, especially at elevated temperatures). It took 70 years to understand[11,12] that the anomalous Hall effect is due to the spin-orbit interaction.

This effect can also be seen in nonmagnetic conductors, where the spin polarization is created by application of a magnetic field. The spin-related anomalous effect can be separated from the much larger ordinary Hall effect by magnetic resonance of the conduction electrons, which results in a resonant change of the Hall voltage.[13] Nonequilibrium spin polarization produced either by optical means or by spin injection, should also result in an anomalous Hall voltage. Such an experiment was recently reported by Miah and co-workers[14] on GaAs illuminated by circularly polarized light.

- *Electric current induced by* $\nabla\mathrm{x}\boldsymbol{P}$

The term $\delta\nabla\mathrm{x}\boldsymbol{P}$ in Eq. (5) describes an electric current induced by an inhomogeneous spin density (now referred to as the inverse SHE). It can also be regarded as the diffusive counterpart of the anomalous Hall effect.

A way to measure this current under the conditions of optical spin orientation was proposed in Ref. 3. The circularly polarized exciting light is absorbed in a thin layer near the surface of the sample. As a consequence, the optically created electron spin density is inhomogeneous, however $\nabla\mathrm{x}\boldsymbol{P} = 0$, since $\boldsymbol{P}$ is perpendicular to the surface and it varies in the same direction. By applying a magnetic field parallel to the surface, one can create a parallel component of $\boldsymbol{P}$, thus inducing a non-zero $\nabla\mathrm{x}\boldsymbol{P}$ and the corresponding surface electric current (or voltage).

This effect was observed by Bakun *et al.*,[4] providing the first experimental observation of the inverse SHE, see Fig. 2. In a later publication, Tkachuk *et al.*[5] observed very clear manifestations of the nuclear magnetic resonance in the surface current induced by $\nabla\mathrm{x}\boldsymbol{P}$.

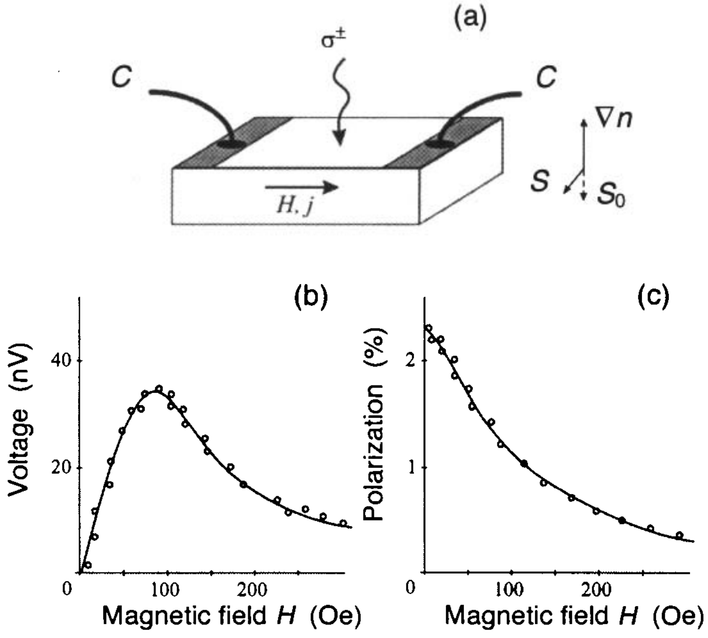


Figure 2. First experimental observation of the inverse spin Hall effect. (a) The experimental setup; (b) voltage measured between the contacts *C* *vs.* magnetic field *H*; (c) measured degree of circular polarization of luminescence, equal to the normal component of the average electron spin, as a function of magnetic field. The solid line in (b) is calculated using the results in (c).

- *Current-induced spin accumulation or the spin Hall effect*

The terms $\beta n \varepsilon_{ijk} E_k$ and its diffusive counterpart $\delta \varepsilon_{ijk} \partial n/\partial x_k$ describe the spin Hall effect: an electrical current induces a transverse spin current, resulting in spin accumulation near the sample boundaries.[1] This phenomenon was observed experimentally only in recent years[6,7] and has attracted widespread interest.

Spin accumulation obtains from solving Eq. (7) in the steady state ($\partial P/\partial t = 0$) and using Eq. (6) for the spin current. Since the spin polarization will be proportional to the electric field, to the first order in E terms involving the EP product can be neglected.

We take the electric field along the x-axis and look at what happens near the boundary $y = 0$ of a wide sample situated at $y > 0$. The boundary condition corresponds to vanishing of the normal to the boundary component of the spin current, $q_{yj} = 0$.

The solution of the diffusion equation $D d^2\mathbf{P}/dy^2 = \mathbf{P}/\tau_S$ with the boundary conditions at $y = 0$ that follow from Eq. (6), $dP_x/dy = 0$, $dP_y/dy = 0$, $dP_z/dy = \beta nE/D$, gives the result:[1]

$$P_z(y) = P_z(0)\exp(-y/L_S), \quad P_z(0) = -\beta n E L_S/D, \quad P_x = P_y = 0\,, \tag{8}$$

where $L_S = (D\tau_S)^{1/2}$ is the spin diffusion length.

Thus the current-induced spin accumulation exists in thin layers (the *spin layers*) near the sample boundaries. The width of the spin layer is given by the spin-diffusion length L_S, which is typically on the order of 1 μm. The polarization within the spin layer is proportional to the driving current, and the signs of spin polarization at the opposing boundaries are opposite. These predictions were confirmed by experiments in Refs. 6, 7, and many others.

It should be stressed that all these phenomena are closely related and have their common origin in the coupling between spin and charge currents given by Eqs. (3) and (4). Any mechanism that produces the anomalous Hall effect will also lead to the spin Hall effect and *vice versa*. It is remarkable that there is a single dimensionless parameter γ that governs the resulting physics. The calculation of this parameter should be the objective of a microscopic theory.

- *The degree of polarization in the spin layer*

Using Eq. (8), the *degree* of spin polarization in the spin layer, $\mathrm{P} = P_z(0)/n$, can be rewritten as

$$\mathrm{P} = \gamma \frac{v_D}{v_F}\sqrt{\frac{3\tau_S}{\tau_P}} \tag{9}$$

where v_F is the Fermi velocity, τ_P is the momentum relaxation time, we have introduced the electron drift velocity $v_D = \mu E$ and used the conventional expression for the diffusion coefficient of a degenerate 3D electron gas $D = v_F^2\tau_P/3$. (For 2D electrons, the factor 1/3 becomes 1/2, whereas if the electrons are nondegenerate, v_F should be replaced by the thermal velocity.)

In materials with inversion symmetry, like Si, where both the spin-charge coupling and spin relaxation via the Elliott–Yafet mechanism are due to spin asymmetry in scattering by impurities, the strength of the spin-orbit interaction cancels out in Eq. (9), since $\tau_S \sim \gamma^{-2}$.

Thus, the most optimistic estimate for the degree of polarization within the spin layer is $P \sim v_D/v_F$. In semiconductors, this ratio may in principle be on the order of unity. In the absence of inversion symmetry, usually the Dyakonov–Perel' mechanism makes the spin relaxation time considerably shorter, which is unfavorable for appreciable spin accumulation. So far, the experimentally observed polarization is on the order of 1%.

- *The validity of the approach based on the diffusion equation*

The diffusion equation is valid when the scale of spatial variation of concentration (in our case, of spin polarization density) is large compared to the mean free path $l = v_F\tau_P$. The variation of $\boldsymbol{P}$ occurs over the spin diffusion length, so the condition $L_S >> l$ should be satisfied. Since $L_S \sim l(\tau_S/\tau_P)^{1/2}$, this condition can be equivalently re-written as $\tau_S >> \tau_P$.

Thus, if the spin relaxation time becomes comparable to the momentum relaxation time (in the so-called "clean limit", when the spin band splitting is greater than $\hbar/\tau_P$), the diffusion equation approach breaks down. The diffusion equation still can be derived for spatial scales much greater than l, but it will be of no help for the problem at hand, because neither this equation, nor the boundary conditions for the spin current can any longer be used to study spin accumulation. Surface spin effects will occur on distances less than l from the boundaries and will crucially depend on the properties of the interfaces (e.g. flat or rough interface, *etc.*). To understand what happens near the boundaries, one must address the quantum-mechanical problem of electrons reflecting from the boundary in the presence of electric field and spin-orbit interaction.

5. Swapping of spin currents: additional terms in Eq. (4)

Pure symmetry considerations allow for additional terms in one of our basic equations, Eq. (4). Namely, it is possible to complement the right hand side of Eq. (4) for the spin current by additional terms of the type: $q_{ji}^{(0)}$ (note the transposition of the indices i and j!) and $\delta_{ij}q_{kk}^{(0)}$ (the sum over repeating indices is assumed) with some new coefficients proportional to the spin-orbit interaction.[1,2] This means that the spin-orbit interaction will transform the spin currents, for example, turn $q_{xy}^{(0)}$ into q_{yx}, *i.e.* the directions of spin and flow will be interchanged: flow of the y component of spin in the x direction will induce the flow of the x component in the y direction.

This *swapping* of spin currents should lead to new transport effects.[15] For example, suppose that the spins are aligned in the z direction. An electrical current in the x direction will be accompanied by the spin current $q_{xz}^{(0)}$, which will induce the spin current q_{zx}. Now spins oriented along x will flow to the boundaries located in the z direction. This will lead to accumulation of x-oriented spins at these boundaries, resulting in a slight current-induced rotation of the boundary spin polarization around the y-axis.

The physical origin of spin currents swapping will be discussed in the next section.

- *Effects of spin asymmetry in electron scattering*

Mott has shown[15,16] that spin-orbit interaction results in an asymmetric scattering of polarized electrons. If a polarized electron beam hits a target, it will deviate in a direction depending on the sign of polarization (similar to a spinning tennis ball in air). This effect is used at high-energy facilities to measure the polarization of electron beams (the Mott detectors).

The scattering of electrons by a charged center is schematically depicted in Fig. 3. The most important element for us is the magnetic field ***B*** existing in the electron's moving frame and seen by the electron spin. This field is perpendicular to the plane of the electron trajectory and has opposite signs for electrons moving to the right and to the left of the charged center. The Zeeman energy of the electron spin in this field is, in fact, the spin-orbit interaction.

Simply *looking* at Fig. 3, one can make the following observations:

- *Electron spin rotates.* If the electron spin is not exactly perpendicular to the trajectory plane, it will precess around ***B*** during the time of collision. The angle of spin rotation during an individual collision depends on the impact parameter and on the orientation of the trajectory plane with respect to spin. This precession is at the origin of the Elliott–Yafet mechanism of spin relaxation.

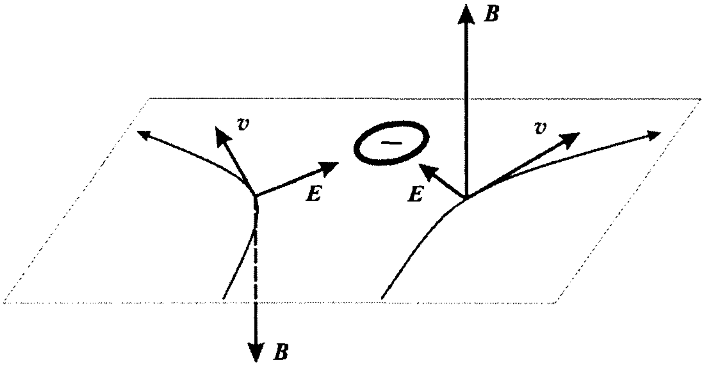


Figure 3. Schematics of electron scattering by a negatively charged center. The electron spin sees a magnetic field $\boldsymbol{B} \sim \boldsymbol{v} \times \boldsymbol{E}$ perpendicular to the plane of the electron trajectory. Note that this magnetic field has opposite directions for electrons scattered to the left and to the right.

- *The scattering angle depends on spin.* The magnetic field $\boldsymbol{B}$ in Fig. 3 is inhomogeneous in space because the electric field $\boldsymbol{E}$ is nonuniform and also because the velocity $\boldsymbol{v}$ changes along the trajectory. For these reasons, there is a spin-dependent force (proportional to the gradient of the Zeeman energy), which acts on the electron. As a consequence, a left-right asymmetry appears in scattering of electrons with a given spin. This is the Mott effect, or *skew scattering*, resulting, among other things, in the anomalous Hall effect. If the incoming electrons are *not* polarized, the same spin asymmetry will result in separation of spin-up and spin-down electrons. Spin-up electrons will mostly go to the right, while spin-down electrons will go to the left, so that a spin current in the direction perpendicular to the incoming flux will appear (the SHE).
- *Spin rotation is correlated with scattering.* As seen from Fig. 3, the spin rotation around the field $\boldsymbol{B}$ is correlated with scattering. If the spin on the right trajectory (corresponding to scattering to the right) is rotated clockwise, then the spin on the left trajectory (scattered to the left) is rotated counter-clockwise. The existences of this correlation and its consequences for spin transport[17] have been never discussed previously.

Let us see what happens if the incoming beam (along the x axis) is polarized along the y-axis in the trajectory plane, *i.e.* characterized by a spin current q_{xy}. After scattering, the electrons going to the right will have a spin component along the x-axis, while the electrons going to the left will have an x component of the opposite sign! This means that scattering will partly transform the initial spin current q_{xy} to q_{yx}. Similarly, q_{xx} will transform to $-q_{yy}$.

Such an analysis shows that during the scattering process the initial spin current $q_{ij}^{(0)}$ generates a new spin current q_{ij} according to the rule $(q_{ji}^{(0)} - \delta_{ij}q_{kk}^{(0)}) \to q_{ij}$. Thus, the correlation between spin rotation and the direction of scattering gives a physical reason for an additional term in Eq. (4), which should be modified as follows:[17]

$$q_{ij} = q_{ij}^{(0)} - \gamma\varepsilon_{ijk}q_k^{(0)} + \kappa(q_{ji}^{(0)} - \delta_{ij}q_{kk}^{(0)}) . \qquad (10)$$

The last term with the dimensionless coefficient κ describes the swapping effect. The effect of skew scattering (γ) appears only beyond the Born approximation. By contrast, the swapping of spin currents (κ) is a more robust effect: it exists already in the Born approximation. The swapping coefficient κ was calculated in Ref. 17.

In fact, spin-dependent effects in scattering are described by *three* different cross-sections.[17] The first of them describes skew scattering, leading to the SHE, the second one describes the swapping effect, and the third one is responsible for the Elliott-Yafet spin relaxation. These cross-sections were introduced a long time ago by Mott and Massey[16] to analyze the results of a single scattering event in atomic physics.

A possible way to observe the swapping effect[17] is presented in Fig. 4. The primary spin current $q_{yy}^{(0)}$ is produced by spin injection in a semiconductor sample through a ferromagnetic contact. Swapping will result in the appearance of transverse spin currents $q_{xx} = q_{zz} = \kappa q_{yy}^{(0)}$. Those secondary currents will lead to an excess spin polarization near the lateral boundaries of the semiconductor sample, which could be detected by optical means. At the top face there will be a polarization $P_z < 0$ and at the bottom face $P_z > 0$ (similarly $P_x < 0$ at the front face and $P_x > 0$ at the back face). The accumulation of spins polarized perpendicular to the surface distinguishes this manifestation of swapping from the SHE, where the accumulated spins are parallel to the surface.

6. What is it good for, practically?

Thus far, nobody knows. My personal opinion is that the spin Hall effect is one of the many physical phenomena (like, for example, the Nernst effect or the Shubnikov-de Haas effect) that have no practical applications. Still, it is good to know and understand them, because such phenomena may help determine some material parameters and, more importantly, because based on this knowledge we may be able to understand other effects of eventual practical importance.

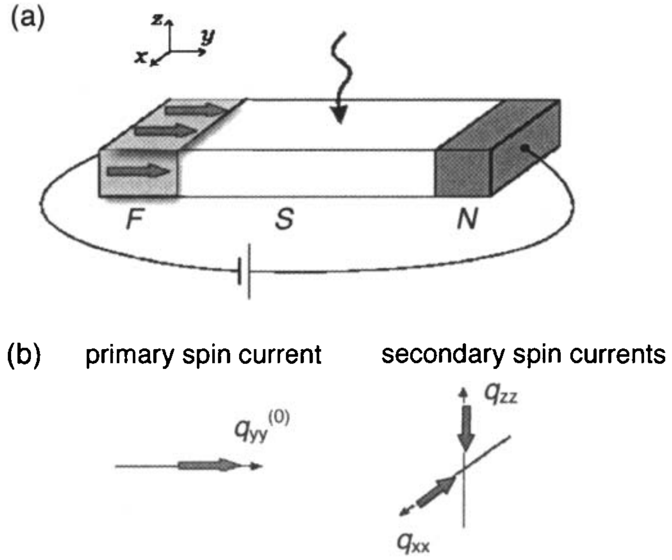


Figure 4. Schematics of a proposed experiment revealing the swapping of spin currents. (a) The $q_{yy}^{(0)}$ spin current is electrically injected in a semiconductor (*S*) through a ferromagnetic contact (*F*) with a magnetization along the *y* axis. The wavy arrow symbolizes optical detection of the *z*–component of spin polarization near the surface; (b) the swapping effect transforms the primary spin current $q_{yy}^{(0)}$ into q_{xx} and q_{zz}. This should lead to the appearance of excess spin polarization at the lateral boundaries of the sample ($P_z < 0$ at the top face and $P_z > 0$ on the bottom face) that may be detected optically.

However, virtually every paper devoted to this subject starts with the standard mantra: "implementing electrical manipulation of the spin degree of freedom is an essential element for the emerging field of spintronics", or something to this effect. Whether such manipulation (with useful consequences) will ever be possible remains to be seen, but to date there are no indications of it.

Another issue that is sometimes raised in the current literature, and especially in articles designed for the general public, and which I consider to be pure hype, is the "dissipationless spin transport for spin-based electronics". It is proclaimed that dissipationless intrinsic spin currents open the shining perspective of reducing the power consumption in future spin-based computers, *etc.*

These statements should be taken with a big grain of salt. First, *all* spin currents by themselves, independently of their microscopic mechanism, are dissipationless. This is simply a consequence of the fact, mentioned in the Section 2, that the spin current, unlike the charge current, does not change sign under time inversion. Indeed, the spin current is described by the "spin Hall mobility" β, which is a dissipationless kinetic coefficient. Second, the existence of dissipationless spin currents does not mean that one can save energy, because the spin current is induced (with low efficiency) by a charge current, which *does* involve dissipation. The normal Hall current is dissipationless too. This does not mean that we may reduce energy dissipation in future computers by utilizing Hall currents!

7. Conclusions

The spin Hall effect is a new transport phenomenon, predicted a long time ago but observed only in recent years. It was experimentally studied in 3D and 2D semiconductors samples. [6,7,16-21] The inverse spin Hall effect has been seen in semiconductors,[4,5,19] as well as in metals.[22,23] Finally, it is important that these effects are observable not only at cryogenic, but also at room temperature.[23,24] However, the number of experimental papers is still about two orders of magnitude less than the number of theoretical ones. At present, it is difficult to predict whether this effect will have any practical applications, as many people believe, or will remain a fundamental research tool for studying spin interactions in solids.

References

1. M. I. Dyakonov and V. I. Perel', "Possibility of orienting electron spins with current," *Pis'ma Zh. Eksp. Teor. Fiz.* **13**, 657 (1971) [*JETP Lett.* **13**, 467 (1971)].
2. M. I. Dyakonov and V. I. Perel', "Current-induced spin orientation of electrons in semiconductors," *Phys. Lett. A* **35**, 459 (1971)
3. N. S. Averkiev and M. I. Dyakonov, "Current due to inhomogeneity of the spin orientation of electrons in a semiconductors," *Fiz. Tekh. Poluprovodn.* **17**, 629 (1983) [*Sov. Phys. Semicond.* **17**, 393 (1983)].
4. A. A. Bakun, B. P. Zakharchenya, A. A. Rogachev, M. N. Tkachuk, and V. G. Fleisher, "Observation of a surface photocurrent caused by optical orientation of electrons in a semiconductor," *Pis'ma Zh. Eksp. Teor. Fiz.* **40**, 464 (1984) [*JETP Lett.* **40**, 1293 (1984)].
5. M. N. Tkachuk, B. P. Zakharchenya, and V. G. Fleisher, "Resonant photovoltaic effect in the NMR of nuclei in a semiconductor lattice," *Pis'ma Zh. Eksp. Teor. Fiz.*, **44**, 47 (1986) [*JETP Lett.* **44**, 59 (1986)].
6. Y. K. Kato, R. C. Myers, A.C. Gossard, and D. D. Awschalom, "Observation of the spin Hall effect in semiconductors," *Science* **306**, 1910 (2004).
7. J. Wunderlich, B. Kaestner, J. Sinova, and T. Jungwirth, "Experimental observation of the spin Hall effect in a two-dimensional spin-orbit coupled semiconductor system," *Phys. Rev. Lett.* **94**, 047204 (2005).
8. J. E. Hirsch, "Spin Hall effect," *Phys. Rev. Lett.* **83**, 1834 (1999).
9. M. I. Dyakonov and A. V. Khaetskii, "Spin Hall effect," chapter 8 in: M. I. Dyakonov, ed., *Spin Physics in Semiconductors*, Berlin: Springer, 2008.
10. M. I. Dyakonov, "Magnetoresistance due to edge spin accumulation," *Phys. Rev. Lett.* **99**, 126601 (2007).
11. J. Smit, "Magnetoresistance of ferromagnetic metals and alloys at low temperature," *Physica*, **17**, 612 (1951); "The spontaneous Hall effect in ferromagnetics - I," *ibid.* **21**, 877 (1955); "The spontaneous Hall effect in ferromagnetics - II," *ibid.* **24**, 29 (1958).
12. R. Karplus and J. M. Luttinger, "Hall effect in ferromagnetics," *Phys. Rev.* **95**, 1154 (1954).
13. J. N. Chazalviel and I. Solomon, "Experimental evidence of anomalous Hall effect in a nonmagnetic semiconductor," *Phys. Rev. Lett.* **29**, 1676 (1972).
14. M. I. Miah, "Observation of the anomalous Hall effect in GaAs," *J. Phys. D: Appl. Phys.* **40**, 1659 (2007).
15. N. F. Mott, "Scattering of electrons by gold," *Proc. Roy. Soc. A* **124**, 425 (1929).
16. N. F. Mott and H. S. W. Massey, *The Theory of Atomic Collisions*, 3rd. ed., Oxford: Clarendon Press, 1965.
17. M. B. Lifshits and M. I. Dyakonov, "Swapping of spin currents," arXiv: 0905.4409 (to be published).

18. K. Nomura, J. Wunderlich, J. Sinova, B. Kaestner, A. H. MacDonald, and T. Jungwirth, "Edge-spin accumulation in semiconductor two-dimensional hole gases," *Phys. Rev. B* **72**, 245330 (2005).
19. H. Zhao, E. J. Loren, H. M. van Driel, and A. L. Smirl, "Coherence control of Hall charge and spin currents," *Phys. Rev. Lett.* **96**, 246601 (2006).
20. V. Sih, W. H. Lau, R. C. Myers, V. R. Horowitz, A. C. Gossard, and D. D. Awschalom, "Generating spin currents in semiconductors with the spin Hall effect," *Phys. Rev. Lett.* **97**, 096605 (2006).
21. V. Sih, R. C. Myers, Y. K. Kato, W. H. Lau, A. C. Gossard, and D. D. Awschalom, "Spatial imaging of the spin Hall effect and current-induced polarization in two-dimensional electron gases," *Nature Phys.* **1**, 31 (2005).
22. S. O. Valenzuela and M. Tinkham, "Direct electronic measurement of the spin Hall effect," *Nature* **442**, 176 (2006).
23. T. Kimura, Y. Otani, T. Sato, S. Takahashi, and S. Maekawa, "Room temperature reversible spin Hall effect," *Phys. Rev. Lett.* **98**, 156601 (2007).
24. N. P. Stern, S. Ghosh, G. Xiang, M. Zhu, N. Samarth, and D. D. Awschalom, "Current-induced polarization and the spin Hall effect at room temperature," *Phys. Rev. Lett.* **97**, 126603 (2006).

This page intentionally left blank

Can Biology Provide Creative Solutions for Next-Generation Memory Devices?

W. E. van den Berg and S. A. Kushner
Dept. of Psychiatry, Erasmus University Medical Center
Rotterdam, The Netherlands

1. Introduction

With a weight of only 1350 grams, the brain is considered to be the most complex, sophisticated, and efficiently built machine ever known. For the 100 billion neurons within the brain to function, it requires only the power level of a 60-watt light bulb. Shaped by an exceptionally long history of adaptive evolutionary forces, the brain was designed to maximize reproductive fitness.

The most remarkable feature of the brain is its ability to optimize its performance in ever-changing circumstances. It may seem that some parts of the brain are indeed highly pre-programmed to allow for processing of sensory input, motor control, and natural responses.[1] However, recent literature reveals plasticity mechanisms capable of altering the constitution of the brain beyond genetic predetermination.

While details are still being worked out, somehow these alterations encode our experiences. However, we also know that our memory is fallible and it is certainly not always literal. Does this mean that information storage by the brain is not as sophisticated as we think? The brain receives inputs from several sensory systems of the body, and stores the information accordingly. In that sense, memory is a reconstruction of facts and experiences on the basis of the way they were stored, not necessarily as they actually had occurred.[2]

It is considered one of the biggest challenges neuroscientists face at present, and progress is only made slowly, but step by step we are getting closer to unraveling the neural representation of a memory: how the brain encodes information.

2. How is the brain designed

Every vertebrate brain can be divided into three broad zones: the hindbrain, midbrain, and forebrain. The forebrain is related to purposeful, voluntary behavior and is considered to play a role in problem solving. The midbrain is involved in maintaining wakefulness and processing visual and auditory reflexes, whereas the hindbrain controls the basic functions needed for staying alive, such as breathing and maintaining a heart beat. When comparing a mammalian brain to any other

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

vertebrate brain, the expanded cortex in the mammalian brain (which is part of the forebrain) immediately stands out. Among mammals, humans have the most differentiated cortex, and within the cortex our neocortex allows for many of our higher mental functions.

The brain consists of several interconnected systems. Each plays a role in cognitive, emotional, and bodily functions. Of special interest is the limbic system, which is the region that forms the border (*limbus* in Latin) of the inner cortex. This area, consisting of the amygdala, hippocampus, thalamus, and hypothalamus, as well as other regions, is responsible for emotional processing, behavior, and long-term memory. Besides its role as a storage mechanism, the limbic system has a regulatory function, since it has connections to the endocrine system (hormones) and the autonomic nervous system (digestion, respiration).

Within the limbic system, the hippocampus and amygdala are the main structures related to memory formation. The hippocampus plays an important role in long-term memory and spatial navigation. Specifically, episodic memory, defined as the memory for facts and events, and declarative memory, memories that can be explicitly verbalized, are mediated by the hippocampus. Impairment in hippocampal function also results in the inability to transfer information from working memory to the long-term memory capabilities of the hippocampus and cortex.

The amygdala has been identified as the main system in the processing and memory of fear and emotions. It has projections to several areas in the brain responsible for immediate somatic responses to emotional events (fight or flight). It is also interconnected with the hippocampus, providing contextual information to these emotional memories.

As all organs and tissues of the body, the brain is made up of cells: neurons and glia. Neurons are responsible for communication, whereas glial cells have been thought to play a primarily supportive role (however, see Filosa *et al.*[3] for more recent data concerning glial-neuronal communication). Groups of neurons form nuclei and these nuclei form the systems described above. About one million kilometers of interconnected fibers form the infrastructure by which neurons are able to communicate. Neurons, about 100 billion (10^{11}) in total, have the unique ability to exchange information directly. Each neuron is connected to many other neurons, receiving input from around 10,000 others.

Neurons are packed closely together, about 10^5 per mm^3 of brain. Their average diameter is 10 μm, with a surface area of 250,000 μm^2. Neurons have a conduction speed of about 10 m/s, allowing intra-neuronal electrical signals to travel fast compared to typical biological time-scales, albeit quite slowly compared to the processing speed of modern microelectronic devices. In order to communicate, neurons consist of several dendrites (the information "receiving antennas"), a cell body (the biological workshop for the processes within the neuron), and a single axon (information "sending antenna").

The electrical signals in an axon have an all-or-nothing character. Based on the strength of the incoming message, each neuron individually determines if the message should be passed on to the next. If the input is too small, the neuron will not continue to pass on the electrical signal. This means that the relay stops here. Otherwise, the signal continues and could eventually initiate an outcome, such as muscle movement, or inspire a thought. As mentioned earlier, neurons receive inputs from multiple others. When occurring simultaneously, several weak inputs might add up to a potential strong enough to excite the next neuron. This is called convergence; information from several systems in the brain can be integrated into a single signal. On the output side, though neurons usually have only one axon, this axon branches several times before it ends. This allows a single neuron to excite several others. This is called divergence, allowing for quick distribution of information signals.

Communication within the brain is not only about *excitation* of the next neuron. Within local systems, *inhibitory* interneurons play an important role in modifying the information relay. These neurons counterbalance the excitatory outputs of other neurons, allowing for complex feed-forward inhibition and excitation loops.

At an inter-neuronal level, communication occurs most commonly through chemical signals, as the vast majority of neurons are not connected physically (however, gap junctions are present in the mammalian nervous system, which permits rapid and direct current flow between physically-connected neurons). Principal inputs and outputs to a neuron are most often via synapses, located on axon terminals (pre-synaptic) and dendritic branches (post-synaptic). In total, there are approximately 1 quadrillion (10^{15}) synapses, packed together on an average of 10^9 per mm^3 of the brain. Synapses bridge the 40 nm gap between neurons – the so-called synaptic cleft – leading to a conduction delay (across a synapse) of ~5 ms.

Synapses play an important role in determining the immensely plastic capabilities of the brain. New connections form repeatedly, just as unused connections are eliminated. This physiological plasticity is accompanied by axon branching and new synapse formation, both during development and following learning,[4] allowing for experiences and memory to find their way into our biological self.

After formation, frequently used synapses are also strengthened, either for a limited period of time or permanently. These synapses become more prone to pre-synaptic signals, increasing the likelihood to become excited. Time-limited, or short-term synaptic strengthening, is rapidly induced by a potentiation of the neuron. It represents only a functional change: signals are being sent and received more efficiently. In more stable or late-phase synaptic strengthening, synapses undergo physical changes, which depends on protein synthesis. This type of change requires more time and is also less easily induced; multiple potentiations within a short time frame are required.

To summarize, the brain represents an interconnected neural circuit of several systems, which each have their own functions. Communication occurs through electrical signals carried by neurons, which have specific transmission properties. Due to various experiences, these neuronal transmission properties are altered by

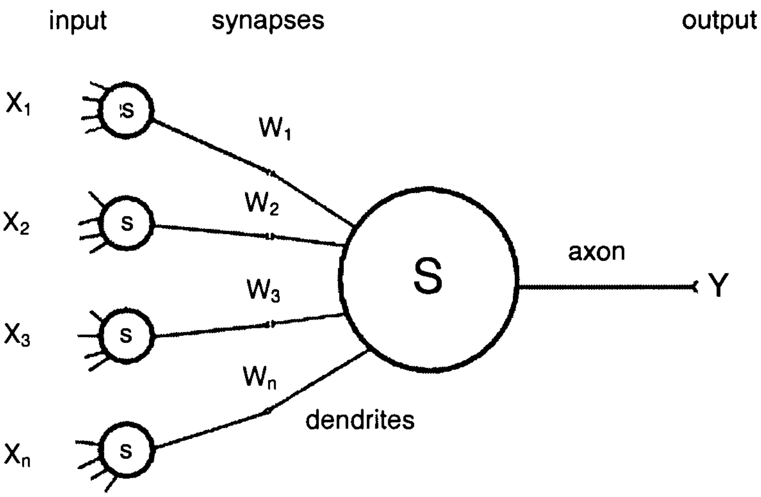


Figure 1. Schematic representation of a convergent neuronal network, where the input signals X_j are multiplied by the corresponding synaptic weights W_j to determine the binary "all or nothing" output signal (first proposed by McCulloch and Pitts[5]).

the formation and elimination of synaptic connections (among other types of alterations). The immense number of neurons, the formation of even more synaptic connections, and the plasticity of the weight of every individual synapse, taken together determine the almost endless amount of differentiated patterns of connections that can be stored in the brain.

3. What does the brain do exceptionally well?

Neuro-biologists, as well as psychologists, psychiatrists, and philosophers, have long debated about the influence of nature (genetic predispositions) as opposed to nurture (experience) on development and behavior. Currently, scientists generally reject the notion that at birth, the brain is a blank slate (*tabula rasa*) or that genes determine everything. However, some complex capabilities of the brain are so efficiently acquired and effortlessly utilized, that they can almost be considered an innate quality. It is these processes that might contribute to the development of next-generation memory devices.

Most humans are particularly skilled at extracting meaningful patterns from complex stimuli. As an example, consider face recognition. Multiple inputs need to be integrated and stored in such a way that the data are easily accessible. To make the problem even more complex, all faces have essentially the same anatomical make-up – they look remarkably similar. Humans seem to have developed a very effective solution for this, though its underlying neural correlates are not yet fully understood.

Differentiation between subtle differences in spatial relations between facial features (eyes, nose, mouth) could therefore be the underlying memory formation mechanism. However, we do know that faces are processed holistically – as a whole – rather than as a collection of individual facial features.[6] This actually makes sense, since a face is effortlessly recognized; even in a crowd and even when you are not expecting to see a person at a given moment or location. In addition, consider the fact that caricatures are easily connected to person A, despite the fact that in reality another individual B, who would never be mistaken for person A, usually looks more like person A than that caricature.

In a recent study, single neurons were found to be selectively activated by strikingly different pictures of given individuals and, in some cases, even letter strings with their names, suggesting that some neurons might encode an abstract representation of an individual.[7] These neurons might represent the downstream endpoint of the integration of multiple inputs, resulting from the convergent organization of that neuronal network.

Indeed, the brain is extremely capable of identifying faces with great ease, while the psychic process that produces the experience of face recognition cannot be consciously described.

The rapid acquisition of language is also an exceptional quality of the brain. The essence of language is, as Chomsky pointed out, its infinity.[8] We are able to create new sentences of any possible length, expressions never before imagined by another brain, and engage in conversations the outcome of which could never have been anticipated beforehand – and our brains will always be able to follow. Finding words and setting up sentences does not require conscious effort – at least, in a person's native language.

Besides our ability to handle complex input stimuli, the brain is also responsible for complex output signals, such as speech production and motor coordination.

Imagine a tennis player who needs to respond to a ball that bounced right on the chalk line, changing its speed and direction. Here again, there is no time for the brain to consciously calculate the exact difference in the trajectory of the ball. The body just responds, with speed and precision. In fact, a frequently used training technique within professional sports intended to achieve mastery over the "perfect movement" (let it be a penalty kick in soccer or a free throw in basketball) is to stop trying to consciously interfere with what the body is doing. Make the movement, and evaluate it afterwards. If an adjustment is needed, do not try to consciously alter your movement or even to address specific muscles, but just "tell" the body that the ball should go more to the right. This way, the player takes advantage of the brain's capabilities to coordinate muscle movement unconsciously in an exceptional manner.

4. Where does the brain fall short?

Besides the aforementioned unmatched qualities of the human brain, there are some operations for which the human brain performs relatively poorly compared to computers.

On the input side, this includes the encoding of facts and figures for organized and accurate retrieval. While information storage capabilities are improving rapidly on computers, the brain seems to fall short compared to the enormous amounts of data that can be stored on even a simple NAND flash memory chip. Also, accurate retrieval is much better handled by modern computing devices, compared to the brain. As mentioned earlier, the brain is under continuous influence of the environment, and is continuously changing. As a result, memory retrieval is always done by a brain that is different from the one that stored the information in the first place. This might account for some of the differences that occur when the human brain retrieves information.

Also, the brain is not very good at explicit structured processing of information, such as doing mathematical calculations. We all experience difficulty when explicitly trying to calculate the mathematical properties of a complex differential equation, whereas a computer is able to do this quite rapidly. The processing power of the brain to explicitly perform mathematical computations lags far behind that of even the simplest digital calculator.

5. How can biology inspire microelectronics engineers?

Looking further into the strengths and weaknesses of the human brain provides some clues as to how the brain might inspire microelectronics engineers in trying to develop next-generation memory devices.

The first observation that emerges is the fact that the brain performs most optimally when we are not consciously thinking. Adjusting your position to catch a fly ball is relatively simple, while explicitly calculating the course of the ball is not. In facial recognition, we have no conscious access to either the way we store or retrieve recognition information, but we are able to recognize a familiar person within fractions of a second.

Evolutionary forces are shaped most strongly by genetic fitness, so traits that increase a person's chance of transmitting successful genes will prevail. Hence the conscious/unconscious balance must be drawn from the competing interests of an organism having the ability to consciously reflect on its choices compared with the loss of processing power because of a more limited unconscious function.

Recent studies support the idea that episodic memory, which requires conscious recollection of the time and place of some personal experience, is particularly characteristic of humans. In contrast, semantic memory, being the simple storage of a fact rather than a personal experience, is within the capacity of many animals.[9] The ability to consciously experience and reflect has emerged within the human brain in order to improve evolutionary fitness, meaning that we seem to have traded the ability to be conscious of our actions, for the ability to be maximally good at them!

So, how is the brain able to learn and detect these patterns? Patterns are stored (and learning occurs) only after intense training with a few *a priori* conditions. First, there is the incredible amount (10^{11}) of parallel hubs of information storage. Second, these hubs are each interconnected with about 10^4 other hubs. As a result of experiences, the hubs and connections continuously update and encode the latest available information into the brain; accordingly, these points of contact between the hubs (synapses) represent the main sites of plasticity.

Competition between neurons seems to play an important role in efficiently handling the plasticity of the connections. Only a fraction – we have shown that during a specific fear conditioning task this fraction is only 25% – of eligible neurons (those that receive sensory input) participate in encoding a particular experience.[10] With an average neuronal loss of 31 million per year and the extensive, though limited capacity, of the brain, the number of neurons involved in a given memory must exist as a balance between the goal of having a stable memory (*e.g.*, without memory loss if just a single neuron dies) and the goal of being able to store a large amount of information (*e.g.*, without needing a large majority of all neurons to participate in encoding every memory).

Within the ability to update, the brain seems to have only a few *a priori* basic rules for plasticity. The system efficiently determines patterns autonomously. In order for neurons to connect and form synapses, two neurons need to engage in a simultaneous and reciprocal process in which the input strength, duration, and frequency determines the stability of each individual synaptic connection.

6. Conclusions

Taken together, experience-dependent plasticity has emerged as the dominant theme of information storage within the brain. It allows for biological encoding of an almost endless amount of different memory traces due to the immense supply of synaptic connections. Neuronal competition plays an important role in memory formation, however, we are only beginning to understand which specific neurons are recruited for a memory trace and why.

It seems that evolutionary forces have endowed our brain with consciousness, allowing us to reflect upon our own actions and experiences. This did, however, come at the expense of lower processing power, which becomes evident when we perform explicit, conscious tasks such as mathematical calculations or organized retrieval of facts. We have yet to discover the specific neurological underpinnings of a memory. However, modern neuroscience, together with genetics and microelectronics-driven advances in biological techniques are providing novel opportunities for understanding the fascinating mechanisms underlying human brain function.

References

1. P. Rakic, "Specification of cerebral cortical areas," *Science* **241**, 170 (1988).
2. D. L. Schacter, "The seven sins of memory. Insights from psychology and cognitive neuroscience," *Amer. Psychol.* **54**, 182 (1999).
3. A. Filosa, S. Paixao, S. D. Honseck, *et al.*, "Neuron-glia communication via EphA4/ephrin-A3 modulates LTP through glial glutamate transport," *Nature Neurosci.* **12**, 1285 (2009).
4. K. C. Martin and E. R. Kandel, "Cell adhesion molecules, CREB, and the formation of new synaptic connections," *Neuron* **17**, 567 (1996).
5. W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biol.* **5**, 115 (1943).
6. J. J. Richler, M. L. Mack, I. Gauthier, and T. J. Palmeri, "Holistic processing of faces happens at a glance," *Vision Res.* **49**, 2856 (2009).
7. R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain," *Nature* **435**, 1102 (2005).
8. N. Chomsky, "A review of B. F. Skinner's verbal behavior," *Language* **35**, 26 (1959).
9. J. LeDoux, *Synaptic Self: How Our Brains Become Who We Are*, New York: Viking Adult, 2002, p. 108.
10. J.-H. Han, S. Kushner, and A. P. Yiu, "Selective erasure of a fear memory," *Science* **323**, 1492 (2009).

Spin Screening of Magnetization Due to Inverse Proximity Effect in Superconducting/Ferromagnetic Bilayers

V. Shelukhin, M. Karpovski, A. Palevski
School of Physics and Astronomy, Tel Aviv University
Tel Aviv 69978, Israel

J. Xia, A. Kapitulnik
Department of Physics, Stanford University
Stanford, California 94305, U.S.A.

A. Tsukernik
The Center for Nanoscience and Nanotechnology, Tel Aviv University
Tel Aviv 69978, Israel

1. Introduction

When ferromagnetic (F) films are deposited on top of superconductors (S) they acquire certain superconducting properties. This is the well-known proximity effect which gives rise to a number of spectacular phenomena:[1,2] superconducting critical current oscillations as a function of ferromagnet thickness, long range triplet superconductivity, *etc.* These effects have been already observed experimentally.[3-5] However, the inverse proximity effect,[6,7] wherein the superconductor inherits ferromagnetic properties, is a subject that so far has not attracted as much experimental attention. Here, we present the transport and optical studies of the inverse proximity effect in superconductor ferromagnetic bilayers. We have measured the magneto-resistance of an array of e-beam lithographically patterned 200 nm thick Pb discs with diameters in the range of 2–5 μm, deposited on top of a 5 nm Ni film. Magneto-optical measurements of the polar Kerr effect using a zero-area-loop Sagnac magnetometer on Pb/Ni and Al/(Co-Pd) proximity-effect bilayers show unambiguous evidence for the "inverse proximity effect," in which the ferromagnet (F) induces a finite magnetization in the superconducting (S) layer.[8]

2. Theory of inverse proximity effect: Reduction of the magnetization due to superconductivity

There are two different mechanisms leading to the reduction of the total magnetization in SF structures with itinerant ferromagnets below the

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

superconductor transition temperature T_C. The first mechanism[6,7] is due to the reduction of the magnetization in the ferromagnet, and the second one is due to the induced magnetization[6,7] in the superconductor in the direction opposite to magnetization in F-layer. The reduction of the magnetization in F-layer is caused by the reduction in the density of states in the ferromagnetic material. This is a very small effect in a strong itinerant ferromagnet, where the exchange energy is much larger than the super-conductor gap.

The induced magnetization in the superconductor is a much larger effect,[6,7] that can completely screen the magnetization of a magnetic layer if the thickness of the latter is smaller than its magnetic length. Below T_C, Cooper pairs of electrons of opposite spin separated by a typical distance ξ_S form the superconducting state. If the magnetization is homogeneous in the ferromagnet (no rotation), then the condensate function penetrating the ferromagnet due to proximity effect is a mixture of the singlet and triplet components with zero projection of the total spin, and can be viewed as electron pairs with opposite spins just like in the case without an F layer. It is clear that when both electrons forming the pair are in the S material, they cannot contribute to the magnetization. Some pairs, however, are distributed in a more complicated manner, namely one electron is in the ferromagnet whereas another one is in the superconductor. The spin of the electron in the F-layer is aligned along the direction of the magnetization for an itinerant ferromagnet; therefore the spin of another electron will have an opposite direction, contributing to a magnetic moment induced in the superconductor. This is, of course, an oversimplified description, but it explains the results obtained by a rigorous calculation in the limit where the thickness of the ferromagnet is smaller than the magnetic length.

At zero temperature and for the ideal SF interface, the magnetization of the ferromagnet is completely screened.[6,7] This spin screening effect should not be confused with the Meissner effect. The latter is an orbital effect that produces a diamagnetic moment in the superconductor that screens the external magnetic field. The induced magnetization in the superconductor due to the spin screening considered in this paper takes place in the SF bilayers even if the magnetization in the ferromagnet does not produce a magnetic field in the superconductor. As an example, consider an infinite plane of magnetic material with the magnetic moment perpendicular to the plane in contact with the plane of the superconductor. The magnetic field B_{IN} inside the material, in the presence of external magnetic field B_{EXT} and magnetization $\mathbf{M}$ in the material, is given by $B_{IN} = B_{EXT} + (1-K)\mu_0\mathbf{M}$ and depends on the value of the demagnetization factor K. For an infinite plane, K is unity. Therefore, in the absence of external magnetic field, the ferromagnetic thin film magnetized in the direction normal to the plane will not produce a magnetic field at any point in real space and thus a Meissner diamagnetic moment will not be induced, whereas a diamagnetic moment due to the spin screening will occur as discussed above.

3. Sample fabrication

Samples for both transport and optical measurements were made of thin films of superconductor material deposited on top of thin films of ferromagnetic material. We used a variety of strong itinerant ferromagnets like Ni, and Pd (3 Å)/Co(9 Å) *n*-period superlattice structure with $n \sim 10$. A strong ferromagnet like Ni has a magnetic length $L_M \sim 1$ nm and it is appropriate for the studies of the induced magnetization. The ferromagnetic layers were deposited onto SiO_2 and/or GaAs substrates by e-beam evaporation in a vacuum deposition system at the pressure of $2x10^{-7}$ torr. The thickness of the magnetic films ranged between 2 and 10 nm. The superconductors, either Al with a large superconductor coherence length ($\xi_S \sim 100$ nm) or Pb ($\xi_S \sim 60$ nm), were deposited *in-situ* on top of the magnetic layers. This procedure guarantees a good quality SF interface. The thickness of the superconductor films was somewhat larger than ξ_S.

For transport measurements, e-beam lithography was used in order to fabricate an array of small superconductor dots on top of the ferromagnetic layer. The lateral dimensions of the dots as well as the separation between them were of the same order as the thickness of the superconductor layers, namely about 100 nm.

4. Transport experiments

We have measured the magnetoresistance of a discontinuous Pb superconductor deposited on the top of a 5 nm Ni film. The Hall effect of the sample above T_C is shown in Fig. 1(a) and the HRSEM micrograph of the sample in Fig. 1(b).

We see from Fig. 1(a) that the Hall effect curve exhibits relatively pronounced hysteresis. From this measurement we conclude that Ni is fully magnetized in the out-of-plane direction. Otherwise the remnant magnetization at zero external field would be smaller that the saturated magnetization at large fields, due to formation of Ni domains, with preferred in-plane direction of magnetization.

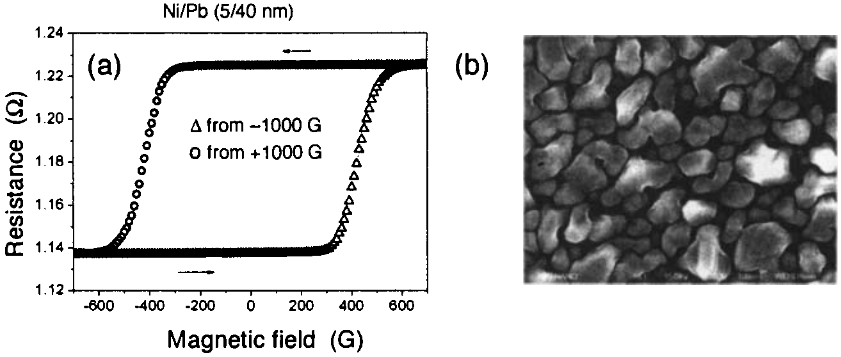


Figure 1. (a) Hall effect of Ni/Pb sample above the transition temperature (9 K), two sweeps from −1 kG to +1 kG and from +1 kG to −1kG; (b) HRSEM image of Pb on top of Ni, size of Pb grains: 100–500 nm.

From the geometry of the Pb layer, as shown in Fig. 1(b), one can expect that the demagnetization factor $K < 1$, and therefore the magnetic field can be induced if the magnetization appears in Pb below T_C.

The magnetoresistance curve of the same bilayer, shown in Fig. 2(a), exhibits relatively strong hysteresis (~10 G) in the range of magnetic fields where the magnetization is reversed. This is expected, since in this range of fields the Ni film produces a magnetic field, *i.e.* $K < 1$ when the domain structure appears. Moreover, a parallel component of the field will also appear and contribute to the increase of the resistance. However, a small hysteresis (~4 G) persists down to zero field, where Ni is fully magnetized – see Fig. 2(b). Note that the direction of hysteresis coincides with the direction of magnetization in the Ni layer. This means that the field induced in Pb has opposite direction to the Ni magnetization, and could be a result of the *magnetization induced in Pb.*

Of course, we cannot make a definitive statement that the observed hysteretic shifts arise from the inverse proximity effect; however, this experiment indicates that this transport method is sensitive enough to measure local magnetization.

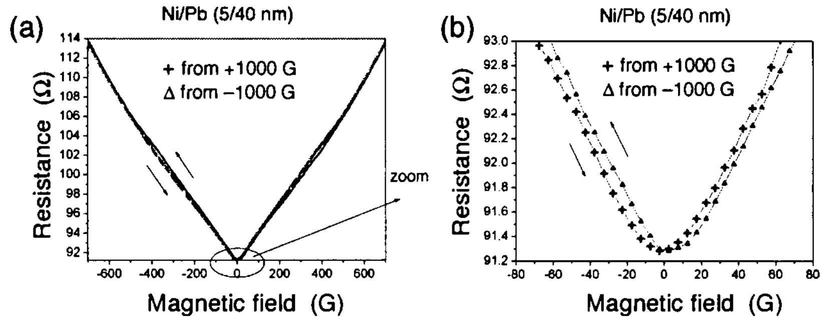


Figure 2. (a) Magnetoresistance curve of Pb/Ni bilayers at 4.2 K; (b) close-up of the data at low fields.

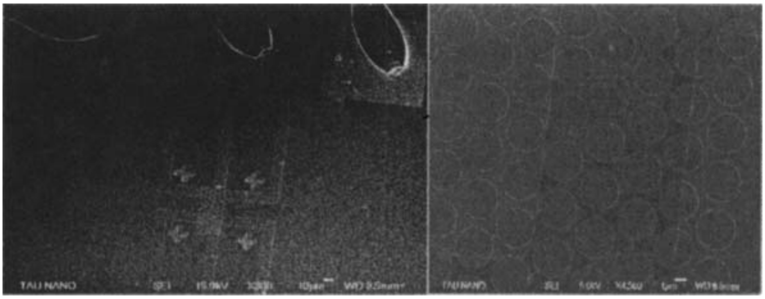


Figure 3. HRSEM image of the array of Pb discs on top of Co/Pd n = 12 superlattice strip.

To demonstrate further the relevance of this method, we utilized the artificially prepared superconductor dots with much smaller demagnetization factor than that of Pb used in the above experiment. We have measured the magnetoresistance of a e-beam lithographically patterned 200 nm thick Pb discs with diameters in the range 2–5 μm, deposited on the top of a 5 nm Ni (or Co/Pd $n = 12$ superlattice) film. The high-resolution SEM images of the samples are shown below in Fig. 3.

After measurements of the Hall effect above the T_C of the superconductor (producing results similar to those in Fig. 1(a)), the samples were cooled below T_C for the induced magnetization experiments. In order to remain sensitive to the induced magnetization, we could not cool below $0.9T_C$, because at lower temperatures the critical magnetic field of the Pb discs becomes too high to perform relevant magnetoresistance measurements.

The magnetoresistance curves of 5 μm and 3 μm diameter Pb discs on top of the Co/Pd $n = 12$ superlattice film are shown in Figs. 4(a) and (b). Both curves exhibit relatively strong anomalous hysteresis of about 3 G and 7 G respectively. The direction of the hysteresis coincides with the direction of magnetization in Co/Pd superlattice, indicating that the field induced in Pb has opposite direction to F layer magnetization, and hence a result of the magnetization induced in Pb.

The effect we observe in the fabricated array of superconducting discs essentially is the same as observed in semi-continuous Pb films of Fig. 2. The hysteresis increases as the diameter of the superconducting Pb discs decreases. This is expected since the demagnetization factor is reduced when the ratio of the height of the disc to its diameter is decreased. It is difficult to measure temperature dependence of the hysteresis, because after transition to superconducting state below T_C the critical magnetic field in the superconducting material becomes too large.

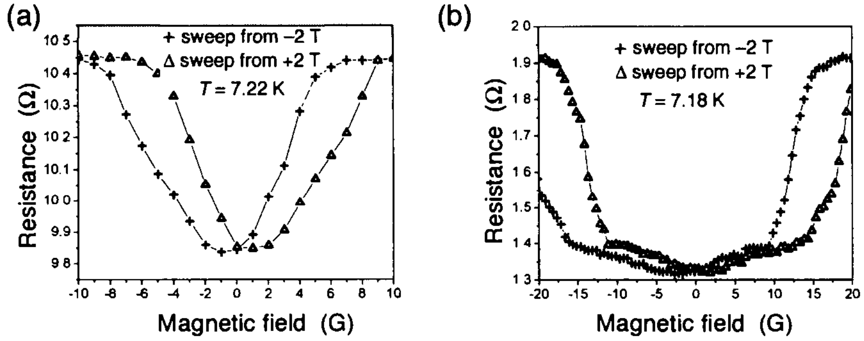


Figure 4. (a) Magnetoresistance of the Pb (5 μm diameter, 200 nm thick discs) on Co/Pd $n = 12$ multilayer at $T_C = 7.22$ K; (b) magnetoresistance of the Pb (3 μm diameter, 200 nm thick discs) Co/Pd $n = 12$ multilayer near the superconducting transition temperature, at $T = 7.18$ K.

5. Optical measurements

Our Ni/Pb and Co-Pd/Al bilayers, with superconductor temperatures $T_C \approx 6.5$ K for Pb/Ni and $T_C \approx 0.6$ K for Al/Pd-Co structures were mounted into the novel apparatus that was built at Stanford University[9] for polar Kerr effect (PKE) measurements, Fig. 5.

Our samples of typically 5x5 mm size were mounted on a copper plate of the cryostat using GE varnish for magneto-optical studies at low temperatures. The system was aligned at room temperature, focusing the beam that emerges out of the quarter-waveplate to a 3-μm spot in the middle of the samples. Polar Kerr effect measurements were performed using a zero-area-loop polarization-Sagnac interferometer at $\lambda = 1550$ nm. The light is reflected from the S top layer of our structures. Typical performance was a shot-noise limited 0.1 μrad/pHz at 10 μW of incident optical power from room temperature down to 0.5 K. The samples were first cooled at zero external fields to about 10 K. At this temperature an external magnetic field of 6 Tesla normal to the film was applied to magnetize the ferromagnetic layer in the direction of the applied magnetic field and after that the field was turned off at the same temperature, well above the superconductor transition. The samples were cooled to $T = 0.5$ K at zero magnetic field and the Kerr effect was measured while the samples were warmed from 0.5 K to the temperatures above T_C. Figure 6 shows the Kerr effect measured on both Pb/Ni and Al/Co-Pd samples. Both graphs clearly show the onset of Kerr angle variation at a temperature that, according to our transport studies, coincides with T_C.

The sign of the Kerr angle detected in our measurement corresponds to the direction of the magnetization in superconductor that is opposite to the direction of magnetization in the ferromagnetic layer. This diamagnetic response is expected for the inverse proximity effect in the diffusive limit.[6]

Pb has much stronger spin-orbit coupling than Al. Nevertheless, the Kerr angle rotation we observe in Al is almost by order of magnitude larger than in Pb. Our estimate of the Kerr angle due to Meissner diamagnetism gives value even for Pb below the detection limit of our apparatus. Therefore, the diamagnetic response observed in our samples must have a spin origin, and it is induced in the superconductor layer by spin screening due to inverse proximity effect.

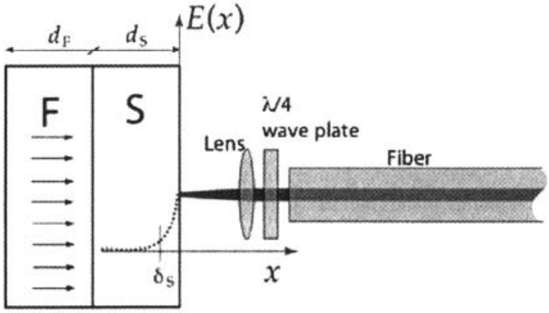


Figure 5. Schematic diagram of the measurement set-up for the polar Kerr effect.

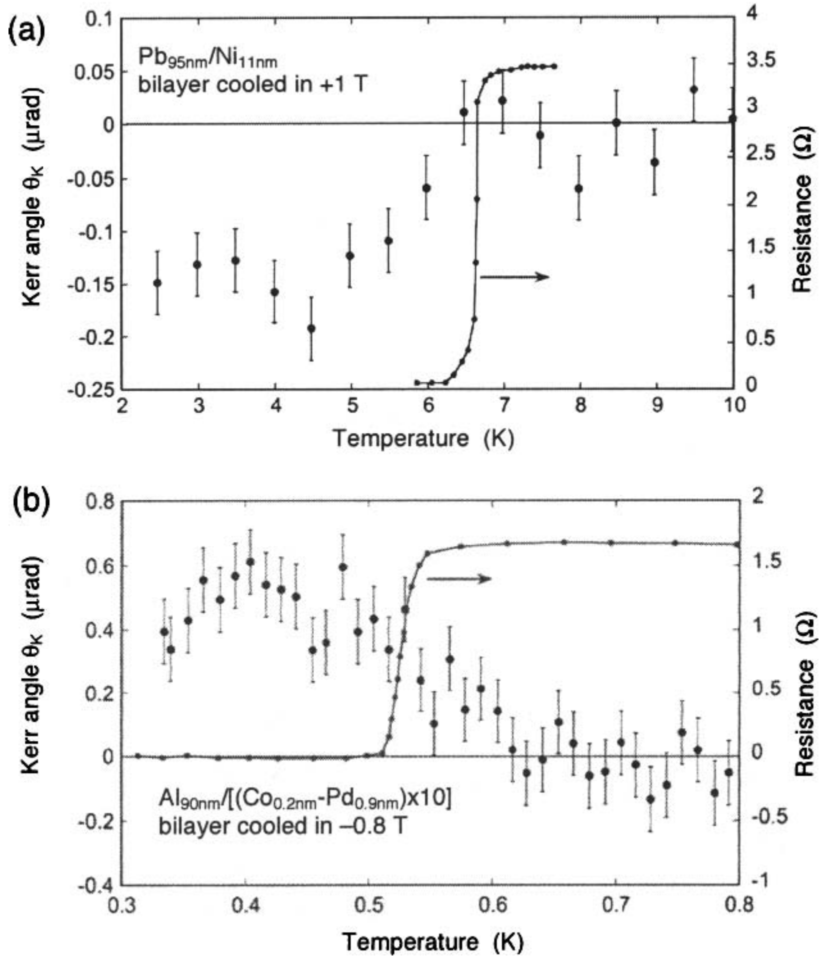


Figure 6. Kerr angle *vs.* temperature (dots) and superconducting transition (line) for: (a) Ni/Pb sample; (b) Co-Pd/Al sample.

5. Conclusions

Inverse proximity effect in SF structures (Pb/Ni and Al/Co-Pd) by transport and optical methods were investigated. The results obtained in both transport and optical studies of the inverse proximity effect strongly indicate that the signals (rotation of the polarization angle of light and the hysteresis) arise from spin-induced magnetization. The increase of the Kerr angle as the temperature is lowered through the superconductor transition temperature follows the theoretically predicted curve and therefore should be considered as direct experimental evidence for the phenomenon called "spin screening" due to proximity effect in S/F bilayers.

Acknowledgments

This work was supported by the US-Israel Binational Science Foundation (BSF) Grant 2006010; NSF NSEC Grant 0425897; Department of Energy Grant DE-AC02-76SF00515.

References

1. A. I. Buzdin, "Proximity effects in superconductor-ferromagnet heterostructures," *Rev. Mod. Phys.* **77**, 935 (2005).
2. F. S. Bergeret, A. F. Volkov, and K. B. Efetov, "Odd triplet superconductivity and related phenomena in superconductor-ferromagnet structures," *Mod. Phys.* **77**, 1321 (2005).
3. V. V. Ryazanov, V. A. Oboznov, A. Yu. Rusanov, A. V. Veretennikov, A. A. Golubov, and J. Aarts, "Coupling of two superconductors through a ferromagnet: evidence of a pi-junction," *Phys. Rev. Lett.* **86**, 2427 (2001).
4. Y. Blum, A. Tsukernik, M. Karpovski, and A. Palevski, "Oscillations of the superconducting critical current in Nb-Cu-Ni-Cu-Nb junctions," *Phys. Rev. Lett.* **89**, 187004 (2002).
5. R. S. Keizer, S. T. B. Goennenwein, T. M. Klapwijk, *et al.*, "A spin triplet supercurrent through the half-metallic ferromagnet CrO_2," *Nature* **439**, 825 (2006).
6. F. S. Bergeret, A. F. Volkov, and K. B. Efetov, "Odd triplet superconductivity and related phenomena in superconductor-ferromagnet structures," *Europhys. Lett.* **66**, 111 (2004).
7. M. Y. Kharitonov, A. F. Volkov, and K. B. Efetov, "Inverse proximity effect in superconductor-ferromagnet bilayer structures," *Phys. Rev. B* **73**, 054511 (2006).
8. Jing Xia, V. Shelukhin, M. Karpovski, A. Kapitulnik, and A. Palevski, "Inverse proximity effect in superconductor-ferromagnet bilayer structures," *Phys. Rev. Lett.* **102**, 087004 (2009)
9. Jing Xia, P. T. Beyersdorf, M. M. Fejer, and A. Kapitulnik, "Modified Sagnac interferometer for high-sensitivity magneto-optic measurements at cryogenic temperatures," *Appl. Phys. Lett.* **89**, 062508 (2006).

Silicon for Spintronic Applications: Strain-Enhanced Valley Splitting

V. Sverdlov, O. Baumgartner, T. Windbacher, and S. Selberherr
Institute for Microelectronics, TU Wien
Gußhausstraße 27-29/E360, A-1040 Wien, Austria

1. Introduction

The rapid increase in computational power and speed of integrated circuits is supported by the incessant downscaling of semiconductor devices. Thanks to constantly introduced innovative changes in the technological processes, the miniaturization of MOSFETs epitomized by Moore's Law successfully continues. The recently introduced 32 nm technology[1] involves improved high-κ gate stack dielectrics integrated with metal gates. These advances, first introduced for the 45 nm MOSFET process technology by Intel,[2] represent some of the major changes in the technological processing since the invention of MOSFETs. Although alternative channel materials with mobility higher than in silicon have already been investigated,[3,4] it is believed that silicon will remain the main channel material for MOSFETs beyond the 32 nm technology node.

With scaling apparently approaching its fundamental limits, the semiconductor industry is facing critical challenges. New engineering solutions and innovative techniques are required to improve CMOS device performance. Strain-induced mobility enhancement, first introduced for the 90 nm technology node, is now routinely used to increase the device performance. With the 4th generation of stressors employed in the 32nm node,[1] strain technology will certainly maintain its key position among engineering solutions for future technology generations. In addition, new device architectures based on multi-gate structures with better electrostatic channel control and reduced short-channel effects are attracting increased attention. A multi-gate MOSFET architecture is expected to be introduced for the 22 nm technology node. Combined with a high-κ dielectric/ metal gate technology and strain engineering, a multi-gate MOSFET appears to be the ultimate device for high-speed operation with excellent channel control, reduced leakage currents, and low power budget. Confining carriers within a thin film reduces the channel dimension in the transverse direction, which further improves gate channel control.

At the same time the search for post-CMOS device concepts has accelerated. In principle, quantum computation promises to open new horizons to approach large-scale computations. Quantum-mechanical properties used in the data structure architecture are speculated to lead to substantial advantages over classical approaches. Although the concept is very attractive, thus far quantum calculations

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

have been demonstrated on a small number of qubits only. Several extremely challenging problems prevent the building of a large quantum computer, and more research is needed. One particular problem is how to increase the quantum state coherence within qubits so that the quantum information is not lost before an acceptable number of operations is performed.

Spin as a degree of freedom is promising for future nanoelectronic and spintronic applications. A recently proposed concept of a racetrack memory[5] is based on the controlled domain wall motion by spin-polarized current in magnetic nanowires. Nonvolatile spin-torque based memory elements combined with gate-all-around MOSFETs are a sound option for a future computer architecture with reduced heat generation.[6]

Silicon, the main element of microelectronics, possesses several properties attractive for spintronic applications. Silicon is composed of nuclei with predominantly zero spin, which greatly reduces fluctuations of local magnetic fields and increases the spin coherence time. Silicon is also characterized by a small spin-orbit coupling. In a recent groundbreaking experiment, coherent spin propagation was demonstrated through an undoped silicon wafer of 350 μm thickness.[7] The experiment was made possible by a unique injection and detection technique of polarized spins delivered through thin ferromagnetic films. Coherent propagation of spins over such long distances makes the fabrication of spin-based switching devices likely already in the near future.

Spin-controlled qubits may be thought of as a basis for upcoming logic gates. However, the conduction band of silicon contains six equivalent valleys, which is potentially a source of increased decoherence.[8] Since the valleys in bulk silicon are degenerate, their quantum numbers are directly competing with the spin quantum numbers. Thus, for successful applications one should be able to lift the valley degeneracy by independently modifying their energies in a controllable way, so that the valley separation can be made larger than the spin Zeeman splitting.

Due to size quantization in (001) oriented Si/SiGe heterostructures, the valley degeneracy is partly lifted. The four valleys with light quantization mass are characterized by higher subband energies, while the two valleys with heavy quantization mass yield the unprimed subband ladder. Yet, within the effective mass approximation usually employed to describe the conduction band, the unprimed subband ladder is doubly degenerate. In reality, this degeneracy is removed due to a slightly non-parabolic dispersion of the conduction band. The actual value of the valley, or, to be more precise, the unprimed subband splitting, is a source of controversy, because it seems to vary significantly depending on the experimental setup. For instance, Shubnikov-de-Haas measurements in an electron system of a thin silicon film grown on a SiGe substrate reveal that the valley splitting is on the order of tens of microvolts.[9] At the same time, recent experiments on the conductivity measurements of point contacts created by confining a quasi-two-dimensional electron system in lateral direction with the help of additional gates deposited on the top of the silicon film, demonstrate a splitting between equivalent valleys larger than the spin splitting.[9]

In this chapter, we will try to resolve this inconsistency with an accurate ***k·p***-based model for the conduction band in silicon. We demonstrate that the parabolic band approximation usually employed for subband structure calculations of confined electrons in silicon inversion layers is not sufficient for ultrathin films. The ***k·p*** model includes strain and is shown to be accurate up to energies of 0.5 eV. This model can therefore be used to describe the subband structure in thin silicon films, where the subband quantization energy may reach a hundred meV.

We first describe the subband structure in a thin unstrained silicon film and we demonstrate that the peculiarities of the subband dispersion obtained within the two-band ***k·p*** model result in a linear dependence of the valley splitting on the magnetic field. Furthermore, we show that a large valley splitting is observed in experiments on conduction quantization through a quantum point contact in the [110] direction, but the splitting is suppressed in [100] point contacts. Finally, we demonstrate that the valley splitting is considerably enhanced in films strained in [110] direction.

Our main result can be summarized as follows: a large valley splitting in the confined electron system can be induced by a shear strain component applied to the system. Since strain is routinely used by the semiconductor industry to enhance MOSFET performance, it opens a way to introduce a splitting between the unprimed subbands in a controllable way at little additional cost. By a direct analogy to the magnetic field that controls the spin splitting, shear strain allows one to tune the valley splitting independently and efficiently, thus making silicon attractive to spintronic and "valleytronic"[10] applications.

2. Accurate model for the conduction band in silicon

To describe the conduction band of silicon close to the minimum located at $k_0 = 0.15(2\pi/a)$ away from the X-point, we follow the classic approach of Refs. 11 and 12. The two conduction bands Δ_1 and Δ_2 are degenerate at the X-point and must be included on an equal footing to describe the dispersion close to the X-point minimum. More distant bands separated by larger gaps are treated with the second order ***k·p*** perturbation theory,[12] which results in the following two-band ***k·p*** Hamiltonian:

$$H = \left(\frac{\hbar^2 k_z^2}{2m_l} + \frac{\hbar^2 (k_x^2 + k_y^2)}{2m_t} \right) I + \left(D\varepsilon_{xy} - \frac{\hbar^2 k_x k_y}{M} \right) \sigma_x + \frac{\hbar^2 k_z k_0}{m_l} \sigma_z \,, \qquad (1)$$

where σ_x and σ_z are the Pauli matrixes, I is the 2x2 unity matrix, m_t and m_l are the transverse and longitudinal effective masses respectively, ε_{xy} denotes the shear strain component, $M^{-1} \approx m_t^{-1} - m_0^{-1}$, and $D = 14$ eV is the shear strain deformation potential.[9] This is the only form of the Hamiltonian in the vicinity of the X-point allowed by symmetry considerations.[12] The two-band Hamiltonian (1) results in the following dispersion relations:

$$E = \frac{\hbar^2 k_z^2}{2m_l} + \frac{\hbar^2(k_x^2+k_y^2)}{2m_t} \pm \left(\frac{\hbar^2 k_0^2}{m_l}\right)\sqrt{\left(\frac{k_z}{k_0}\right)^2+\eta^2}\,, \tag{2}$$

where the negative sign corresponds to the lowest conduction band, and

$$\eta = \frac{m_l(D\varepsilon_{xy} - \hbar^2 k_x k_y / M)}{\hbar^2 k_0^2}.$$

We point out that the energy E and k_z are counted from the X-point.

The ***k·p*** model accurately describes the dispersion relation up to energies of about 0.5 eV.[13] It also reproduces correctly the conduction band dispersion under shear strain.[14] Therefore, the ***k·p*** Hamiltonian (1) can be used to describe the subband structure in thin silicon films and inversion layers.

3. Subband dispersion from the *k·p* theory

For (001) silicon films, the confinement potential gives an additional contribution $U(z)I$ to the Hamiltonian (1). In the effective mass approximation described by (1) with the coefficient in front of σ_x set to zero, the confining potential $U(z)$ is known to quantize the six equivalent valleys of the conduction band of bulk silicon into the four-fold degenerate primed and the two-fold degenerate unprimed subband ladders. In ultrathin films, the unprimed ladder is predominantly occupied and hence the more important. The term with σ_x in (1) couples the two lowest conduction bands and lifts the two-fold degeneracy of the unprimed subband ladder. The additional unprimed subband splitting, or the valley splitting, can be extracted from the Shubnikov-de-Haas oscillations and is typically in the order of a few tens μeV.[9] However, the valley splitting is significantly enhanced in a laterally confined two-dimensional electron gas (2DEG).[9] The valley splitting is usually addressed by introducing a phenomenological intervalley coupling coefficient at the silicon interface.[15] Here we investigate the valley splitting based on the two-band ***k·p*** model (1) without introducing any additional parameters.

We approximate the confining potential of an ultra-thin silicon film by a square well potential with infinite potential walls. This is sufficient for the purpose of analyzing the valley splitting in a quasi-2DEG due to interband coupling. The generalization to include a self-consistent potential is straightforward, though numerically involved.[16] Because of the two-band Hamiltonian, the wavefunction Ψ is a spinor with the two components |0> and |1>. For a wavefunction with a spatial dependence of the form $\exp(ik_z z)$, the coefficients A_0 and A_1 and of the spinor components are related via the Schrödinger equation $H\Psi = E(k_z)\Psi$. For a particular energy E there exist four solutions κ_i ($i = 1, \ldots, 4$) for k_z of the dispersion relation (2), whereas the spatial dependence of a spinor component α is in the form $\sum A_\alpha^{i}\exp(i\kappa_i z)$. The four coefficients are determined by the boundary conditions that both spinor components are zero at the two film interfaces. This leads to the following equations for $k_{1,2} = \kappa_{1,2}/k_0$:

$$\tan\left(k_1\frac{k_0 t}{2}\right)=\frac{k_2}{\sqrt{k_2^2+\eta^2}\pm\eta}\,\frac{\sqrt{k_1^2+\eta^2}\pm\eta}{k_1}\tan\left(k_2\frac{k_0 t}{2}\right). \tag{3}$$

The value of

$$k_2=\sqrt{k_1^2+4-4\sqrt{k_1^2+\eta^2}} \tag{4}$$

becomes imaginary at high η values. Then the trigonometric functions in (3) have to be replaced by the hyperbolic ones. Special care must be taken to choose the correct branch of $(k_1^2+\eta^2)^{1/2}$ in (4): the sign of $(k_1^2+\eta^2)^{1/2}$ must be alternated after the argument becomes zero.

For arbitrary parameters, Eq. (3) can only be solved numerically. We present results of the numerical solution in the next section. However, to gain an insight into the results it is sufficient to analyze the solutions for small strain.

Introducing $y_n = (k_1 - k_2)/2$, Eq. (3) can be written in the form:[17]

$$\sin(y_n k_0 t)=\pm\frac{\eta y_n\sin\left(\sqrt{\frac{1-\eta^2-y_n^2}{1-y_n^2}}\,k_0 t\right)}{\sqrt{(1-y_n^2)(1-\eta^2-y_n^2)}}\,. \tag{5}$$

Applying a perturbation approach to Eq. (5) with η as the small parameter, we obtain the following dispersion relation for the unprimed subbands n:

$$E_n^{\pm}=\frac{\hbar^2}{2m_l}\left(\frac{\pi n}{t}\right)^2+\hbar^2\frac{k_x^2+k_y^2}{2m_t}\pm\left(\frac{\pi n}{k_0 t}\right)^2\frac{\left|D\varepsilon_{xy}-\frac{\hbar^2 k_x k_y}{M}\right|}{k_0 t\left|1-\left(\pi n/k_0 t\right)^2\right|}\sin(k_0 t)\,. \tag{6}$$

Equation (6) demonstrates that the unprimed subbands are not necessarily degenerate and degeneracy is preserved only when the shear strain is zero and either $k_x = 0$ or $k_y = 0$.

4. Splitting of unprimed subbands in magnetic field

For zero shear strain the Landau levels in an orthogonal magnetic field B are found from Eq. (6) by using the Bohr-Sommerfeld quantization conditions:

$$E_m^{(1,2)}=\hbar\omega_C\left(m+\frac{1}{2}\right)\frac{\pi}{4\arctan\left(\sqrt{m_{(1,2)}/m_{(2,1)}}\right)}, \tag{7}$$

where

$$m_{(1,2)} = \left(\frac{1}{m_t} \pm \frac{1}{M} \left(\frac{\pi n}{k_0 t} \right)^2 \frac{\sin(k_0 t)}{k_0 t \left| 1 - \left(\pi n / k_0 t \right)^2 \right|} \right)^{-1} , \tag{8}$$

and $\omega_C = eB/(m_1 m_2)^{1/2}c$ is the cyclotron frequency, e is the electron charge, and c is the speed of light. According to (7), the difference $|E_m^{(1)} - E_m^{(2)}|$ varies linearly with the magnetic field.

In Shubnikov-de-Haas experiments there will now occur two sets of resistance oscillations with slightly different periods. Because of the small difference between the masses, the difference in periods will also be small. However, at the Fermi level, the quantum number m, which is proportional to the ratio of the Fermi energy to the cyclotron frequency, is typically very large and may lead to a splitting of several hundred μeV. The difference in the periods can be interpreted as an appearance of an additional energy shift between the equivalent unprimed valleys. Most importantly, the shift is linear in the magnetic field. The linear dependence of the intervalley splitting on the magnetic field will be also observed in the presence of a small intrinsic constant valley splitting, as long as this splitting is much smaller than the Fermi energy. This splitting is possible due to a remaining shear strain and/or conduction band nonparabolicity, which is not accounted for in the two-band $\boldsymbol{k}\cdot\boldsymbol{p}$ theory and which is usually several tens of μeV, thus much smaller than the Fermi energy. For a 10 nm thick silicon film grown on SiGe it follows from (7) and (8) that the valley splitting can be several tens of μeV in a magnetic field of 1 T, which is consistent with the experimental observations.[9]

5. Splitting between the unprimed subbands in a point contact

Now let us consider a point contact in [110] direction realized by confining an electron system of a thin silicon film laterally by depleting the area under additional gates. Without strain, the low-energy effective Hamiltonian in the point contact can be written as:

$$H_{(1,2)} = \frac{\hbar^2 k_x'^2}{2m_{(2,1)}} + \frac{\hbar^2 k_y'^2}{2m_{(1,2)}} + \frac{1}{2}\kappa x'^2 + V_B \quad , \tag{9}$$

where the primed variables are along the [1, 1, 0] and [1, –1, 0] axes, the effective masses are determined by (8), κ is the spring constant of the point contact confinement potential $V(x') = \kappa x'^2/2$ in [1, –1, 0] direction, and V_B is a gate voltage dependent conduction band shift in the point contact.[18] The dispersion relation of propagating modes within the point contact can be written as:

$$E_p^{(1,2)} = \frac{\hbar^2 k_x'^2}{2m_{(2,1)}} + \hbar\omega_{(1,2)}(p + \frac{1}{2}) + V_B \quad , \tag{10}$$

where $\omega_{1,2}^2 = \kappa/m_{1,2}$. Since the energy minima of any two propagating modes with the same p are separated, they are resolved in a conductance measurement through the point contact as two distinct steps. The valley splitting is $\Delta E_p = \hbar|\omega_1 - \omega_2|$. The difference in the effective masses (8) and, correspondingly, the valley splitting can be significantly enhanced by reducing the effective thickness t of the quasi-2DEG, which is usually the case in a gated electron system, when the inversion layer is formed.

In a [100] oriented point contact without strain, the effective Hamiltonian is

$$H^{\pm} = \hbar^2 \frac{k_x^2 + k_y^2}{2m_t} \pm \left(\frac{\pi n}{k_0 t}\right)^2 \frac{\left|\frac{\hbar^2 k_x k_y}{M}\right| \sin(k_0 t)}{k_0 t \left| 1 - (\pi n / k_0 t)^2 \right|} + \frac{\kappa}{2} x^2 .$$

Due to the symmetry with respect to k_y the subband minima in a point contact are always degenerate. For this reason the valley splitting in [100] oriented point contacts is considerably reduced.

6. Strain-enhanced valley splitting

It follows from (6) that shear strain induces a valley splitting linear in strain for small shear strain values:[17]

$$\Delta E_n = 2\left(\frac{\pi n}{k_0 t}\right)^2 \frac{D\varepsilon_{xy}}{k_0 t \left| 1 - (\pi n / k_0 t)^2 \right|} \sin(k_0 t) .$$

The subband splitting is inversely proportional to the film thickness to the third power, and thus can be quite large in thin films.

One can also evaluate the maximum subband splitting in the $\eta \rightarrow \infty$ limit. Here, it follows from (2) that the band dispersion becomes parabolic again around the minimum located exactly at the X-point. The subband quantization energies are determined by the usual quantization conditions, which results in the subband splitting

$$\Delta E_n = \left(\frac{\hbar \pi}{t}\right)^2 \frac{2n-1}{2m_t} .$$

For practically relevant intermediate strain values a numerical solution of Eq. (3) is required.

7. Numerical results

In order to analyze the subband structure in (001) oriented thin silicon films we first approximate the film potential by the square well potential with infinite potential walls. Although not exact, this is a good approximation for thin films. To obtain the values for the subband splitting and effective masses for an arbitrary strain value, Eq. (3) can be solved numerically. Alternatively, the eigenvalues can be found by solving the equations obtained by discretizing the Hamiltonian (1) with the added confinement potential $IU(z)$ and replacing k_z by $-i\partial/\partial z$. The latter method is more general, because it allows the inclusion of a confinement potential of arbitrary form, making self-consistent calculations possible.

We implemented both numerical routines and obtained equivalent results for a square well potential with infinite walls.

The splitting between the unprimed subbands with the same quantum number n normalized to the ground subband energy in unstrained films for a film of thickness t = 3.2 nm is shown in Fig. 1 as function of shear strain. The dependence is not monotonic and strongly depends on the subband number. Even for the ground subbands with $n = 1$, the splitting is comparable to the subband energy at large strain values. The subband splitting increases rapidly as the film thickness is decreased, as demonstrated in Fig. 2. For ultrathin body films the splitting can reach a value comparable to k_BT already at moderate strain values.

Results shown in Fig. 2 demonstrate that the valley splitting can be effectively controlled by adjusting the shear strain and modifying the effective thickness t of the electron system. Uniaxial stress along [110] channel direction, which induces shear strain, is already used by industry to enhance MOSFET performance. Therefore, its application to control valley splitting does not require expensive technological modifications.

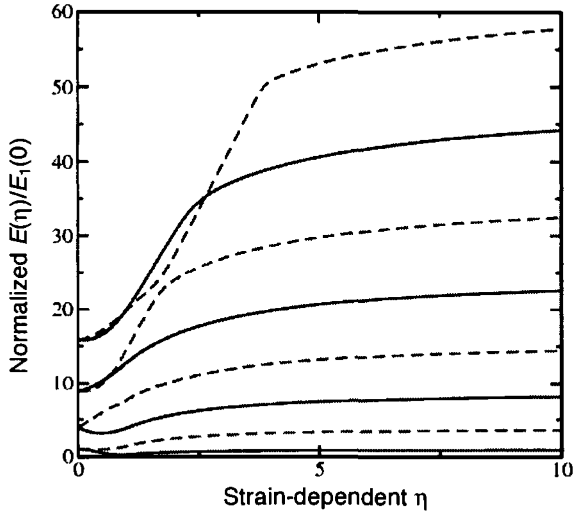


Figure 1. Calculated energies of unprimed subbands in a 3.2 nm thick silicon film as a function of shear strain. The subband energies values are normalized to the energy of the ground subband in an unstrained film. In the unstrained film, the subbands are two-fold degenerate. Strain induces the splitting between the subbands with the same quantum number (shown by solid and dashed lines). The value of the splitting may change sign. The value η = 1 corresponds to the shear strain value ε = 0.016.

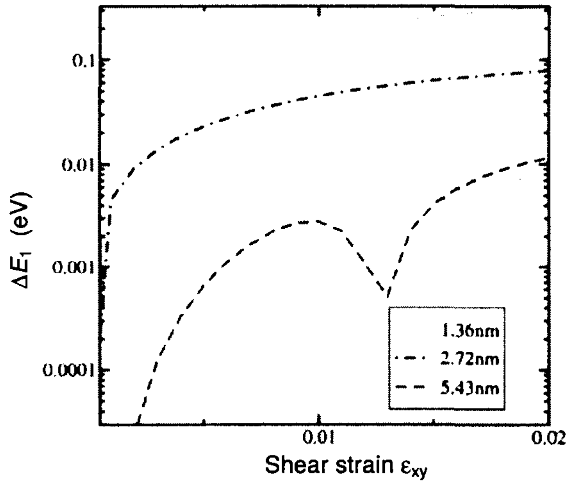


Figure 2. Shear strain-induced splitting of the ground subbands for several film thicknesses. In ultrathin films the splitting exceeds k_BT already for moderate stress.

8. Conclusions

We have analyzed the structure of the unprimed subbands in (001) silicon thin films by using an appropriate ***k·p*** model and demonstrated that the two-fold degeneracy of the unprimed subbands can be lifted, leading to the so-called valley splitting. In an external magnetic field, the model predicts the value of splitting to be proportional to the field. Furthermore, the model predicts that the valley splitting in a <110>-oriented point contact can be larger than the splitting due to the magnetic field because of stronger lateral confinement in a point contact. Finally, the valley splitting can be successfully controlled by shear strain. This makes silicon promising for spintronic and "valleytronic" applications.

Acknowledgments

The authors thank Franz Schanovsky for useful discussions. Financial support from the Austrian Science Fund FWF through the Projects P19997-N14, I79-N16, and SFB IR-ON 2509 is gratefully acknowledged.

References

1. S. Natarajan, M. Armstrong, M. Bost, *et al.*, "A 32 nm logic technology featuring 2nd-generation high-κ + metal-gate transistors, enhanced channel strain and 0.171 μm^2 SRAM cell size in a 291 Mb array," *Tech. Digest IEDM* (2008), pp. 941–943.
2. K. Mistry, C. Allen, C. Auth, *et al.*, "A 45 nm logic technology with high-κ + metal gate transistors, strained Si, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging," *Tech. Digest IEDM* (2007), pp. 247–250.
3. M. K Hudait, G. Dewey, S. Datta, *et al.*, "Heterogeneous integration of enhancement mode $In_{0.7}Ga_{0.3}As$ quantum well transistor on silicon substrate using thin composite buffer architecture for high-speed and low-voltage (0.5V) logic applications," *Tech. Digest IEDM* (2007), pp. 625–628.
4. R. Chau, "Challenges and opportunities of emerging nanotechnology for future VLSI nanoelectronics," *Proc. ISDRS* (2007), p. 3.
5. S. S. P. Parkin, M. Hayashi, and L. Thomas, "Magnetic domain-wall racetrack memory," *Science* **320**, 190 (2008).
6. T. Endoh and T. Hanyu, "Impact of spintronic devices for future nano silicon based LSI," in *Abstracts 5th Intern. School Conf. Spintronics Quantum Inform. Technol. (SPINTECH-V)* (Krakow, 2009), p. 95.
7. I. Appelbaum, B. Huang, and D. J. Monsma, "Electronic measurement and control of spin transport in silicon," *Nature* **447**, 295 (2007).
8. P. Zhang and M. W. Wu, "Spin diffusion in Si/SiGe quantum wells: Spin relaxation in the absence of Dyakonov-Perel' relaxation mechanism," *Phys. Rev. B* **79**, 075303 (2009).
9. S. Goswami, K.A. Slinker, M. Friesen, *et al.*, "Controllable valley splitting in silicon quantum devices," *Nature Phys.* **3**, 41 (2007).
10. A. Rycerz, J. Tworzydlo, and C. W. J. Beenakker, "Valley filter and valley valve in graphene," *Nature Phys.* **3**, 172 (2007).
11. J. C. Hensel, H. Hasegawa, and M. Nakayama, "Cyclotron resonance in uniaxially stressed silicon," *Phys. Rev.* **138**, A225 (1965).
12. G. L. Bir and G. E. Pikus, *Symmetry and Strain-Induced Effects in Semiconductors*, New York: Wiley, 1974.
13. V. Sverdlov, O. Baumgartner, T. Windbacher, *et al.*, "Impact of confinement of semiconductor and band engineering on future device performance," *Proc. 215th ECS Meeting, Silicon-on-Insulator Technol. Dev.* (2009), pp. 15–26.
14. V. Sverdlov, G. Karlowatz, S. Dhar, *et al.*, "Two-band $k{\cdot}p$ model for the conduction band in silicon: Impact of strain and confinement on band structure and mobility," *Solid State Electronics* **52**, 1563 (2008).
15. M. Friessen, S. Chutia, C. Tahan, *et al.*, "Valley splitting theory of SiGe/Si/SiGe quantum wells," *Phys. Rev. B* **75**, 115318 (2007).

16. O. Baumgartner, M. Karner, V. Sverdlov, *et al.*, "Numerical quadrature of the subband distribution functions in strained silicon UTB devices," *Proc. 13th Intern. Workshop Comput. Electronics* (2009), pp. 53–56.
17. V. Sverdlov and S. Selberherr, "Electron subband structure and controlled valley splitting in silicon thin-body SOI FETs: Two-band $k{\cdot}p$ theory and beyond," *Solid State Electronics* **52**, 1861 (2008).
18. B. van Wees, H. van Houten, and C. Beenakker, "Quantized conductance of point contacts in a two-dimensional electron gas," *Phys. Rev. Lett.* **60**, 848 (1988).

This page intentionally left blank

Graphene-Based Terahertz Devices: Concepts and Characteristics

V. Ryzhii, M. Ryzhii, A. Satou, and N. Ryabova
CNEL, University of Aizu, Aizu-Wakamatsu 965-8580, Japan

T. Otsuji
RIEC, Tohoku University, Sendai 980-8577, Japan

V. Mitin
Dept. of Electrical Engineering, SUNY–Buffalo, Buffalo 14260, U.S.A.

F. T. Vasko
Institute of Semiconductor Physics NAS, Kiev 03028, Ukraine

A. A. Dubinov and V. Y. Aleshkin
Institute for Physics of Microstructures RAS, Nizhny Novgorod 603950, Russia

M. S. Shur
Dept. of ECSE, Rensselaer Polytechnic Institute, Troy 12180, U.S.A.

1. Introduction

Graphene, *i.e.* a monolayer of carbon atoms densely packed in a two-dimensional honeycomb structure, graphene bilayers, and patterned graphene structures have captured the great interest of researchers and device engineers. Modification of a laterally uniform graphene layer (GL) into an array of nanostrips, which are usually referred to as graphene nanoribbons, results in the transformation of two-dimensional (2D) linear electron and hole gapless energy spectra into sets of one-dimensional (1D) subbands accompanied by the appearance of an energy gap. Graphene bilayers also exhibit the gapless energy spectra, but with the dispersion relations close to parabolic ones. In a transverse electric field, the energy gap opens and the electron and hole dispersion relations are modified.

Due to the features of the electron and hole energy spectra, particularly their massless, neutrino-like form, as well as of the scattering processes, GLs exhibit unique transport properties that open up wide prospects for their applications in future electronics.[1-3] The features of graphene include the following:

- linear dispersion relation for electrons and holes with zero energy gap;

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

- strong anisotropic interband tunneling and high velocities ($v_W \approx 10^8$ cm/s) of electrons and holes, providing their ballistic transport in devices with micrometer sizes;
- formation of electrically-induced *p* and *n* regions as well as *pin* junctions;
- bandgap engineering possibilities due to lateral quantization in nanoribbons and transverse electric field-induced energy gaps in bilayers;
- high electron and hole mobilities (up to $\mu = 2 \text{x} 10^5$ cm^2/V·s at room temperatures and up to $\mu \sim 10^7$ cm^2/V·s at $T < 55$ K);
- exceptional mechanical properties.

These features open up prospects to create high-speed electron devices, in particular, diodes and transistors, and nano-electromechanical devices surpassing those made of the customary semiconductor materials. Owing to the gapless energy spectrum of electrons and holes in GLs and narrow bandgaps in graphene nanoribbons and graphene bilayers, graphene-based structures appears to be also very promising for various terahertz (THz) devices. In this chapter, we review and analyze several proposals of following devices:

- GL lasers with optical pumping and cascade emission of optical phonons; and
- GL injection lasers and tunneling transit-time devices with electrically induced lateral *pin* junctions.

2. Lasers with optically pumped GLs

A laser structure with one or several optically pumped GLs serving as the active media placed inside a Fabry-Perot resonator is shown in Figs. 1(a) and (b). The proposed optical pumping scheme[4] (see also Refs. 5–7) corresponds to that shown in Fig. 1(c). The graphene structure with the two GLs can be fabricated using a Si substrate covered by a 3C-SiC film (on both sides) and subjected to annealing.[8] It assumes that the electrons and holes generated by optical radiation with the photon energy $\hbar\Omega$ emit a cascade of optical phonons and are accumulated in the energy range $E < \hbar\Omega/2 - N_{max}\hbar\omega_0$. Here $\hbar\omega_0 \approx 200$ meV is the GL optical phonon energy and N_{max} is the integer of the ratio $\hbar\Omega/2\hbar\omega_0$, that is, the maximum number of the optical phonons in the cascade. The accumulation of the photogenerated electrons near the bottom of the conduction band and holes near the top of the valence band produces electron and hole distribution functions $f_e(E) = f_h(E) > 1/2$ over a certain range of energies. The result is population inversion and, consequently, the negative absorption coefficient of low-energy photons. To overcome the losses associated with absorption in the substrate, mirrors, and so on, as well as to enhance the pumping efficiency, structures with multiple GLs appear to be attractive. Recently, perfect multiple GL structures (with the number of GLs up to 100) were fabricated using annealing of SiC substrates.[9] A sketch of the laser with such a multiple-GL structure is shown in Fig. 2(a). However, these multiple-GL structures as the active media exhibit the following two features. First, due to the absorption of optical pumping radiation in multiple GLs, the attenuation of this radiation, and consequently, marked difference in the population and the values of the quasi-Fermi energies $E_F^{(k)}$ of different GLs occurs ($1 \leq k \leq K$ is the index and K is the total number of GLs). Second, there is a buffer bottom GL between the SiC substrate and perfect GLs. The density of electrons in the bottom GL is rather high and the electron Fermi energy E_F^B in this GL is large. This can lead to undesirable absorption of THz photons. The electron population in bottom GL and other GLs is schematically shown in Fig. 2(b). An alternative multiple-GL structure without the bottom GL can be fabricated using chemical/mechanical reactions and transferred substrate techniques (chemically etching the substrate and the highly conducting bottom GL or mechanically peeling the upper GLs and transferring them onto a Si or equivalent transparent substrate).

To obtain the lasing condition one needs to find the GL absorption coefficient as a function of the photon frequency. This coefficient is proportional to the real

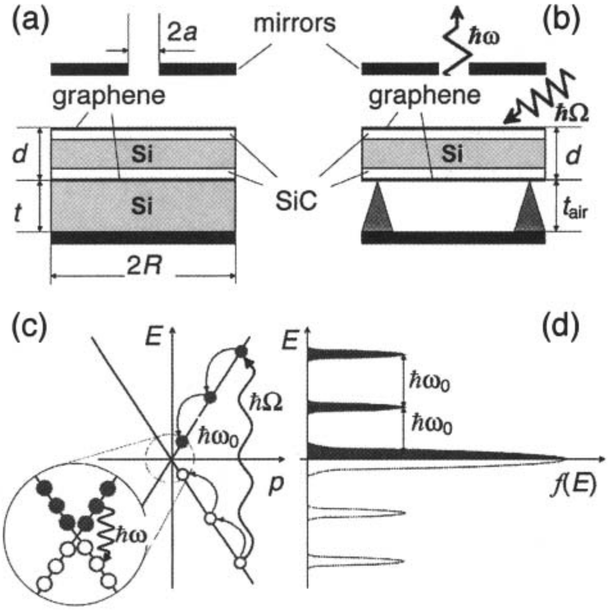


Figure 1. Schematic view of a laser based on a double-GL structure with a Si separation layer (a) and an air separation layer (b), together with the laser pumping scheme (c) and electron and hole distribution functions (d). The jagged arrows in (b) denote the directions of optical pumping radiation and output THz radiation with the photon energies $\hbar\Omega$ and $\hbar\omega$, respectively. Arrows in (c) correspond to the pertinent transitions.

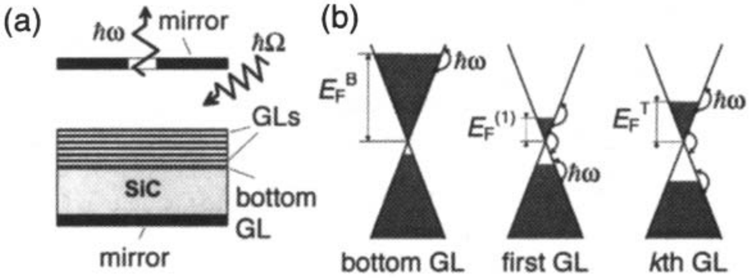


Figure 2. (a) Schematic view of a laser with a multi-graphene-layer (MGL structure; (b) occupied and vacant states in different GLs under optical pumping. Arrows show transitions related to interband emission and intraband absorption of THz photons with energy $\hbar\omega$ (interband optical pumping transitions and intraband relaxation processes of photogenerated electrons and holes are not shown).

part of the dynamic conductivity. By generalizing previously-derived formulae (see, for instance, Ref. 4) to the nonequilibrium electron-hole distributions in the structures under consideration, one obtains the following results for the real part of the dynamic conductivity in the bottom GL (if any) and other GLs, respectively:

$$\begin{aligned}\mathrm{Re}\left\{\sigma_\omega^{\mathrm{B}}\right\} &= \left(\frac{e^2}{4\hbar}\right)\left\{1-\left[1+\exp\left(\frac{\hbar\omega/2-E_{\mathrm{F}}^{\mathrm{B}}}{k_{\mathrm{B}}T}\right)\right]^{-1}-\left[1+\exp\left(\frac{\hbar\omega/2+E_{\mathrm{F}}^{\mathrm{B}}}{k_{\mathrm{B}}T}\right)\right]^{-1}\right\} \\ &+\left(\frac{e^2}{4\hbar}\right)\frac{4k_{\mathrm{B}}T\tau_{\mathrm{B}}}{\pi\hbar\left(1+\omega^2\tau_{\mathrm{B}}^2\right)}\ln\left[1+\exp\left(\frac{E_{\mathrm{F}}^{\mathrm{B}}}{k_{\mathrm{B}}T}\right)\right]\end{aligned} \tag{1}$$

and

$$\mathrm{Re}\left\{\sigma_\omega^{(\mathrm{k})}\right\} = \left(\frac{e^2}{4\hbar}\right)\tanh\left(\frac{\hbar\omega-2E_{\mathrm{F}}^{(\mathrm{k})}}{4k_{\mathrm{B}}T}\right)+\left(\frac{e^2}{4\hbar}\right)\frac{8k_{\mathrm{B}}T\tau}{\pi\hbar\left(1+\omega^2\tau^2\right)}\ln\left[1+\exp\left(\frac{E_{\mathrm{F}}^{(\mathrm{k})}}{k_{\mathrm{B}}T}\right)\right] \tag{2}$$

Here τ_B and τ are the electron and hole momentum relaxation times in the bottom and other GLs, respectively, T is the electron and hole temperature, and $\hbar$ and k_B are the Planck and Boltzmann constants. The quasi-Fermi energy in the kth GL $E_F^{(k)}$ with $k \geq 1$ is mainly determined by the electron (hole) density in this GL, and therefore, by the rate of photogeneration $G_\Omega^{(k)}$ by the optical radiation incident and reflected from the mirror. Using Eq. (2) for $\hbar\omega = \hbar\Omega > 2E_F^{(k)}$, we obtain

$$G_\Omega^{(\mathrm{k})} = \frac{I_\Omega^{(\mathrm{k})}}{\hbar\Omega}\left(\frac{\pi e^2}{\hbar c}\right)\tanh\left(\frac{\hbar\Omega-2E_{\mathrm{F}}^{(\mathrm{k})}}{4k_{\mathrm{B}}T}\right) \simeq \frac{I_\Omega^{(\mathrm{k})}}{\hbar\Omega}\beta\ , \tag{3}$$

where $I_\Omega^{(k)}$ is power density of the optical radiation at the kth GL and $\beta = \pi e^2/\hbar c \approx 0.023$. Considering the attenuation of the optical pumping radiation due to its absorption in each GL, one obtains

$$G_\Omega^{(k)} = \frac{I_\Omega}{\hbar\Omega}\beta\left[(1-\beta)^{K-k} + (1-\beta_B)^2(1-\beta)^{K+k-1}\right]. \qquad (4)$$

Here I_Ω is the intensity of incident optical radiation and $\beta_B = (4\pi/c)\mathrm{Re}\{\sigma_\Omega^B\}$.

To achieve lasing in the multiple-GL structures under consideration, the following condition should be satisfied:

$$Q = \frac{(8\pi/c)\left|\mathrm{Re}\left\{\sigma_\omega^B\right\}(E^B)^2 + \sum_{k=1}^{K}\mathrm{Re}\left\{\sigma_\omega^{(k)}\right\}(E^{(k)})^2\right|}{\left[(1-r_1)E_1^2 + (1-r_2)E_2^2 + (a/R)^2E_1^2 + E_S^2\right]} > 1. \qquad (5)$$

Here, E^B and $E^{(k)}$ are the amplitudes of the THz electric field $E = E(z)$ at the pertinent GL, whereas E_1 and E_2 are the amplitudes of the THz electric field near the bottom and top mirrors, respectively,

$$E_S^2 = (\alpha_S n_S/2)\int_0^{t+d} E^2\,dz \ \text{ or } \ E_S^2 = (\alpha_S n_S/2)\int_0^{t} E^2\,dz,$$

α_S and n_S are the absorption coefficient of THz radiation in the substrate (Si or SiC) and real part of its refractive index, r_1 and r_2 are the reflection coefficients of THz radiation from the mirrors, and a/R is the ration of the diameters of the output hole a and the mirror R. The THz electric field is assumed to be in the GL plane (in the xy plane). In deriving inequality (5), we neglected the finiteness of the multiple-GL thickness (in comparison with t and the THz wavelength) and disregarded the diffraction losses and reflection of THz radiation from the bottom GL.

The spatial distribution of the THz electric field can be found using the following equation:

$$\frac{d^2E}{dz^2} + \frac{\omega^2}{c^2}\kappa_\omega(z)E = 0, \qquad (6)$$

where $\kappa_\omega(z)$ is the permittivity at the THz frequency ω. The spatial dependence of $\kappa_\omega(z)$ reflects a difference in its values in different layers of the resonator (in the air layer, substrate, GL structure, and in the depth of the metal mirrors). The boundary conditions for Eq. (6) correspond to the continuity of E and dE/dz, together with E and $dE/dz \to 0$ when $z \to \pm\infty$.

In the case of double-GL structure (no bottom GL and $K = 2$, see Figs. 1(a) and (b)) with a relatively thick separation layer, one can put $\mathrm{Re}\{\sigma_\omega^B\} = 0$ and, as follows from Eq. (4) due to a smallness of β, $G_\Omega^{(1)} \simeq G_\Omega^{(2)} = I_\Omega\beta/\hbar\Omega$. For the double-GL structure in question, Eq. (5) can be reduced to the following:

$$Q = \frac{(4\pi/c)\left|\mathrm{Re}\left\{\sigma_\omega\right\}\right|\left[(E^{(1)})^2 + (E^{(2)})^2\right]}{\left[(1-r_1)E_1^2 + (1-r_2)E_2^2 + (a/R)^2 E_1^2 + E_S^2\right]}. \tag{7}$$

Taking into account that $E_F^{(1)} \simeq E_F^{(2)} \simeq E_F^T$, the net dynamic conductivity becomes

$$\mathrm{Re}\left\{\sigma_\omega\right\} = \mathrm{Re}\left\{\sigma_\omega^{(1)}\right\} + \mathrm{Re}\left\{\sigma_\omega^{(2)}\right\} \simeq$$

$$\simeq 2\left(\frac{e^2}{4\hbar}\right)\left\{\frac{8k_B T\tau}{\pi\hbar\left(1+\omega^2\tau^2\right)}\ln\left[1+\exp\left(\frac{E_F^T}{k_B T}\right)\right] + \tanh\left(\frac{\hbar\omega - 2E_F^T}{4k_B T}\right)\right\} \tag{8}$$

The first term in the right-hand side of Eq. (8) is associated with the intraband absorption of THz radiation in the GL, whereas the second one is associated with the interband absorption and emission. When the latter is negative, that is when the interband emission prevails the interband absorption, the quantities $\mathrm{Re}\{\sigma_\omega^{(1)}\}$ and $\mathrm{Re}\{\sigma_\omega^{(2)}\}$ exhibit minima as a function of ω. Under sufficiently strong optical pumping when E_F^T is large enough, $\mathrm{Re}\{\sigma_\omega^{(1)}\}$ and $\mathrm{Re}\{\sigma_\omega^{(2)}\}$ can be negative in a certain range of ω. For example, Eq. (8) at $T = 77$ K, $E_F^T = 17$ meV, $\tau = 1$ ps, and $\omega/2\pi = 1.08$ THz yields $\sigma_\omega^{(1)} \simeq \sigma_\omega^{(2)} \simeq -5\times10^7$ cm/s. Figure 3 shows the spatial distribution of the THz electric field and the real part of refractive index for such a laser calculated using Eq. (6) considering real values of $\kappa_\omega(z)$ in different parts of the device. The pertinent quantities Q are shown in the upper right corners of the panels. Here the thicknesses of the Si layers are chosen to be $t = 61$ μm and $d = 39$ μm to maximize the THz electric field modulus just at both GL planes: $E^{(1)} = E^{(2)} = E_{\max}$ – see Fig. 3.

At $r_1 = r_2 = 0.99$, $(a/R) = 0.1$, $\alpha_S \simeq 0.7$ cm^{-1}, $n_S \simeq 3.42$ (for the Si separation layer and the substrate[10]), Eq. (7) provides the maximum value of $Q \simeq 2.16$, which corresponds to the lasing condition. On the other hand, if t and d are not chosen properly, Q can be smaller and even fall below unity.

In the case of the MGL structure, its net thickness is small in comparison with the wavelength of THz radiation (even for $K > 100$), so that the amplitude of the THz electric field is approximately the same in all GLs. When the thickness of the substrate t is chosen to provide the maximum value $E_{\max}$ at the GLs, Eq. (5) can be rewritten as:

$$\begin{aligned} Q &= \frac{(8\pi/c)\left|\mathrm{Re}\{\sigma_\omega^B\} + \sum_{k=1}^{K}\mathrm{Re}\{\sigma_\omega^{(k)}\}\right| E_{\max}^2}{\left[(1-r_1)E_1^2 + (1-r_2)E_2^2 + (a/R)^2 E_1^2 + E_S^2\right]} \\ &= \frac{(8\pi/c)\left|\mathrm{Re}\{\sigma_\omega\}\right| E_{\max}^2}{\left[(1-r_1)E_1^2 + (1-r_2)E_2^2 + (a/R)^2 E_1^2 + E_S^2\right]}, \end{aligned} \tag{9}$$

where

$$\mathrm{Re}\{\sigma_\omega\} = \left(\frac{e^2}{4\hbar}\right)\left\{\frac{4k_BT\tau_B}{\pi\hbar\left(1+\omega^2\tau_B^2\right)}\ln\left[1+\exp\left(\frac{E_F^B}{k_BT}\right)\right]\right.$$
$$\left.+\frac{8k_BT\tau}{\pi\hbar\left(1+\omega^2\tau^2\right)}\sum_{k=1}^{K}\ln\left[1+\exp\left(\frac{E_F^{(k)}}{k_BT}\right)\right]+\sum_{k=1}^{K}\tanh\left(\frac{\hbar\omega-2E_F^{(k)}}{4k_BT}\right)\right\} \qquad (10)$$

is the real part of the net dynamic conductivity of all GLs.

For concreteness, we assume that $E_F^{(k)} \propto [G_\Omega^{(k)}]^\gamma$, where γ is a phenomenological parameter. This leads to

$$E_F^{(k)} = E_F^T\left[(1-\beta)^{K-k}\frac{1+(1-\beta_B)^2(1-\beta)^{2k-1}}{1+(1-\beta_B)^2(1-\beta)^{2K-1}}\right]^\gamma, \qquad (11)$$

where $E_F^T = E_F^{(K)}$ is the quasi-Fermi energy in the topmost GL.

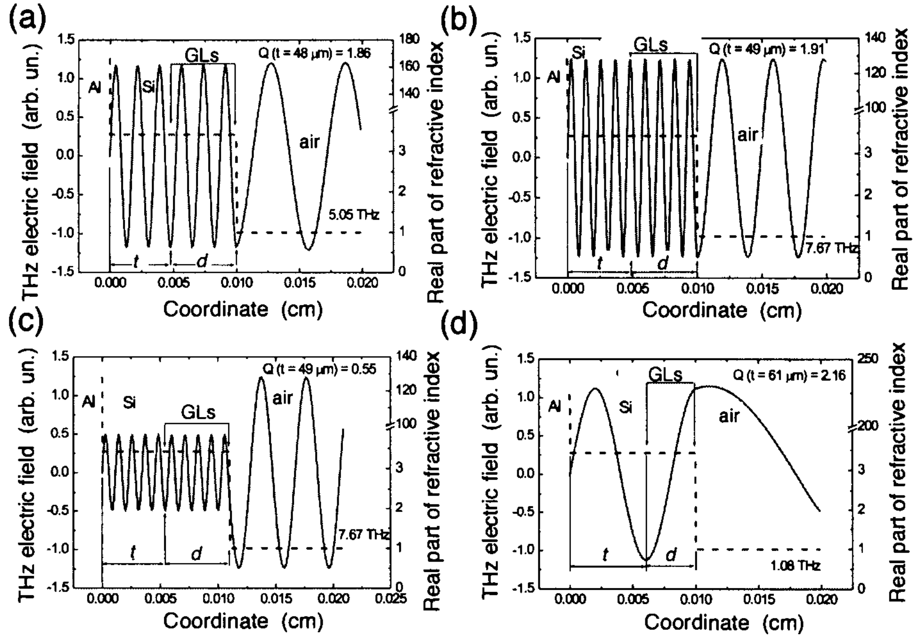


Figure 3. Spatial distributions of THz electric field (solid lines) and real part of refractive index (dashed lines): (a), (b), and (d) correspond to $t + d = 100$ μm, (c) $t + d = 110$ μm.

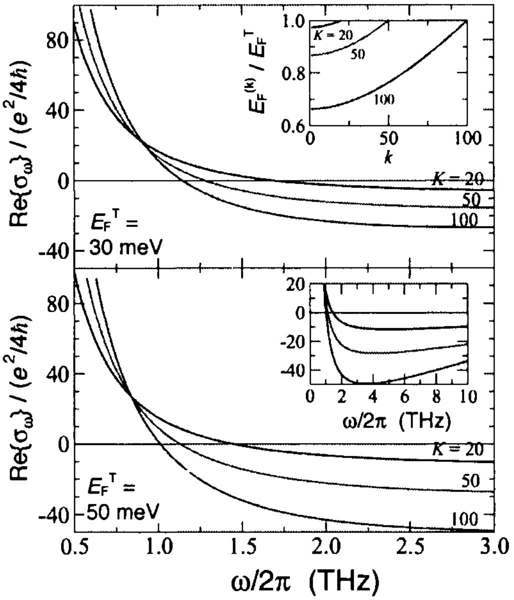


Figure 4. Frequency dependence of the real part of dynamic conductivity $\mathrm{Re}\{\sigma_\omega\}$ normalized by quantity $e^2/4\hbar$ for MGL structures with different number of GLs K at E_F^T = 30 and 50 meV. The inset on upper panel shows how the GL population varies with the GL index k, whereas the inset on lower panel demonstrates the dependences over a wider range of frequencies.

Figure 4 shows the frequency dependence of the real part of the net conductivity $\mathrm{Re}\{\sigma_\omega\}$ normalized by $e^2/4\hbar$ calculated for MGL structures with different numbers of GLs (K = 20, 50, and 100) at different optical pumping intensities (different values of the quasi-Fermi energies in the topmost GL, E_F^T = 30 and 50 meV). As it seen from Fig. 4, an increase in the number of GLs leads to markedly larger absolute value of $\mathrm{Re}\{\sigma_\omega\}$ in the frequency range where it is negative. Figure 5 demonstrates the spatial distribution of the THz electric field inside the Fabry-Perot resonator calculated using Eq. (6) for a laser based on an MGL structure operating at $\omega/2\pi$ = 1.5 THz. The thicknesses of the SiC substrate were chosen as t = 45 μm and t = 80 μm to provide maximum amplitudes of the THz field at the MGL structure. It was assumed that E_F^B = 400 meV, T = 300 K, τ = 10 ps, τ_B = 1 ps, and γ = 0.25. As follows from the calculations utilizing Eq. (11), the quantity $E_F^{(1)}$ is about $0.97E_F^T$, $0.87E_F^T$, and $0.68E_F^T$ at K = 20, 50, and 100, respectively. One can see that $\mathrm{Re}\{\sigma_\omega\}$ becomes negative at $\omega/2\pi > 1$ THz and its absolute value is fairly large. Indeed, if K = 20–100, E_F^T = 30 meV, and $\omega/2\pi$ = 1.5 THz, one obtains $\mathrm{Re}\{\sigma_\omega\} \simeq -(3+10)\mathrm{x}10^8$ cm/s. Assuming, that $r_1 = r_2 = 0.99$, $(a/R) = 0.1$, $\alpha_S \simeq$ 2–4 cm^{-1}, and setting for SiC substrate[11] $n_S \simeq 3$ and t = 45 μm, one obtains $Q \simeq 3.1$–$13.4 \gg 1$. However, in MGL structures will less perfect GL, *i.e.* with shorter momentum relaxation times τ and τ_B, the minimum of $\mathrm{Re}\{\sigma_\omega\}$ is shifted toward higher frequencies and markedly smaller in absolute value.

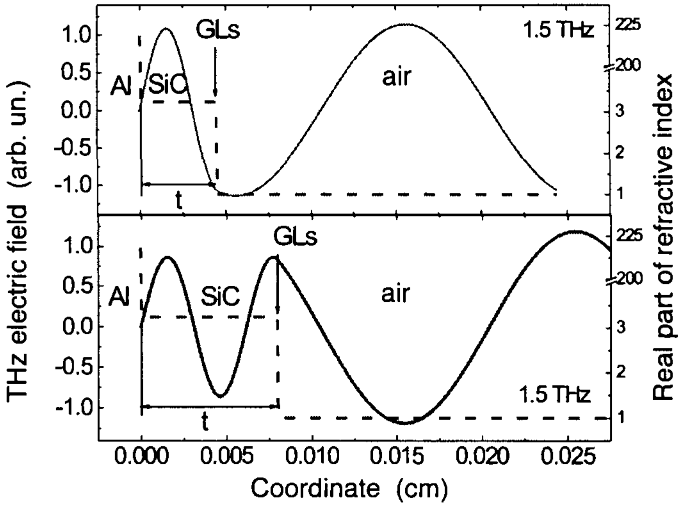


Figure 5. Spatial distribution of the THz electric field (solid line) and real part of refractive index (dashed line) in a laser based on a multiple-GL structure shown in Fig. 2(a) with t = 45 μm (upper panel) and t = 80 μm (lower panel) at $\omega/2\pi$ = 1.5 THz.

Despite stronger intraband absorption characterized by relatively small τ and τ_B, the $Q > 1$ condition can be satisfied in the multiple-GL structures as well, albeit several THz. The GL structures without the bottom GL can provide THz lasing at more liberal conditions.

As follows from the results obtained above, when the quasi-Fermi energy E_F^T in the topmost GL is about 30–50 meV, the conditions for THz lasing at the lower end of the THz range can be satisfied in multiple-GL structures even at the room temperature, particularly if the momentum relaxation times τ and τ_B are sufficiently long. At $T = 300$ K, $E_F^T > 30$ meV corresponds to the electron and hole densities $\sigma_n = \sigma_p \simeq 2\text{x}10^{11}$ cm^{-2} (markedly exceeding the thermal value, which is about $8\text{x}10^{10}$ cm^{-2}). This density can be routinely experimentally achieved in optically pumped GLs. At higher temperatures, the emission of optical phonons can be envisioned as the mechanism of the electron-hole recombination[12] (near the threshold of lasing and not far beyond this threshold). Experimental reports[12-14] have cited densities $\sigma_n = \sigma_p \simeq 2\text{x}10^{11}$ cm^{-2}, achieved at a photogeneration rate $G > 10^{22}$ cm^{-2}s^{-1}. If $\hbar\Omega$ = 120–920 meV, the above photogeneration rate corresponds to the optical power density $I_\Omega > 8$–64 kW/cm^2. Since lowering the temperature leads to a dramatic drop in the rate of the optical-phonon-assisted recombination, the threshold of optical pumping can be decreased by orders of magnitude at lower $T <$ 100 K and be limited by the radiative recombination stimulated by the thermal photons.[6]

3. GL injection laser and tunneling transit-time devices

In gated GL structures, applied gate voltages of different polarities can result in the formation of electron and hole regions in the GLs. In particular, lateral *pn* or *pin* junctions can be electrically induced in GL with split gates.[15] A device structure based on a single-GL with a lateral *pin* junction is shown in Fig. 6(a). When a lateral *pin* junction is forward biased, the electron injection into the *i* and *p* sections and the hole injection into the *i* and *n* sections can lead to population inversion, resulting in a *pin*-based THz injection laser.[16] We can use Eq. (2) with $k = 1$ to calculate the dynamic conductivity for this case. However, the quasi-Fermi energy should be calculated accounting for the carrier injection and the leakage to the source and drain contacts. The electrically induced *pin* junctions can also be realized in the devices like that shown in Fig. 6(a), but with a multiple-GL structure. In this case, the number of effective GLs in which the *p* and *n* sections are electrically induced is limited due to the screening of the transverse electric field (associated with the gate voltages) by the preceding GLs.

The structure with the electrically induced *pin* junction shown in Fig. 6(a) but at reverse bias can also be used in GL tunneling transit-time (G-TUNNETT) devices.[17] The G-TUNNETT structure under consideration and its band diagram (at an applied source-drain voltage V such that the lateral *pin* junction is reverse biased) are schematically shown in Figs. 6(a) and (b), respectively. The operation of the G-TUNNETT in question is associated with the tunneling injection in an electrically induced reverse-biased lateral *pin* junction and the electron and hole transit-time effects in its depleted section.

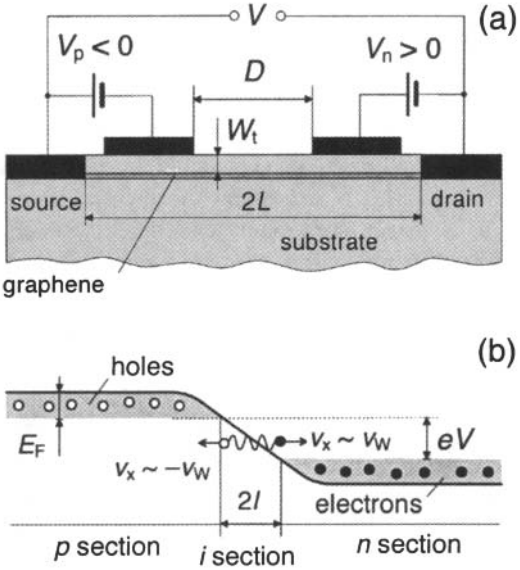


Figure 6. Schematic view of a gated GL structure with electrically induced *pin* junction (a) and the band diagram when the *pin* junction is reverse-biased (b).

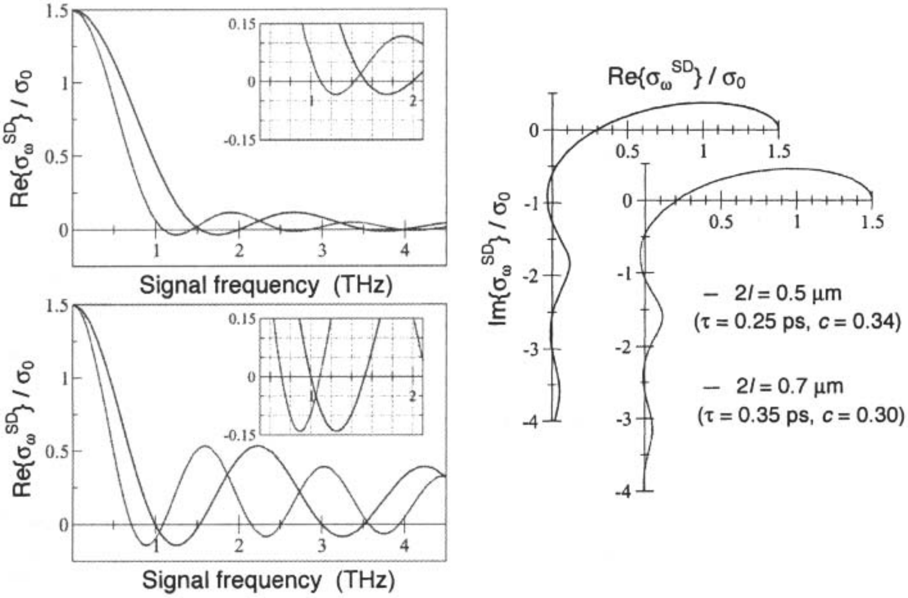


Figure 7. Frequency dependences of real part of the ac conductance calculated for uniform (upper panel) and strongly nonuniform (lower panel) electric-field spatial distributions in the *i* section for different lengths of *i* section: solid lines correspond to $2l = 0.5$ μm ($\tau = 0.25$ ps) and dashed lines correspond to $2l = 0.7$ μm ($\tau = 0.35$ ps). Insets magnify the real and imaginary parts of the ac conductance *vs.* its real part with $\omega\tau$ as a parameter for different lengths of *i* section (right panel).

It is assumed that apart from a dc component of the source drain-voltage $V_0 > 0$, the net source-drain voltage V also contains an ac component $\delta V \exp(-i\omega t)$, where δV and ω are the signal amplitude and frequency, respectively.

The probability of interband tunneling in graphene is a fairly sharp function of the angle between the direction of the electron (hole) motion and the *x*-direction (from the source to the drain). Taking this into account, we disregard the spread in the *x*-component of the velocity of the injected electrons and assume that all the generated electrons and holes propagate in the x-direction with the velocity $v_x \simeq v_W$. Using the formula for the interband tunneling probability,[15] we take into account that the rate of the local tunneling generation $G \propto E^{3/2}$, where $E = E(x)$ is the electric field in the *I*. As a result, for the signal component of the tunneling rate one can obtain $\delta G_\omega = (3G_0/2)(\delta V_\omega/V_0)$, where $G_0 = G_0(x)$ is the dc tunneling rate. Calculating the induced ac terminal (source-drain) current created by the electrons and holes generated in the *i* section and propagating with the velocity $v_x \simeq v_W$ using the Shockley-Ramo theorem, one obtains the following formulae for the ac conductance $\sigma_\omega^{SD} = \delta J_\omega^{SD}/\delta V_\omega$:

$$\sigma_\omega^{SD} = \frac{3\sigma_0}{2}\frac{\sin(\omega\tau)}{\omega\tau}\mathfrak{J}_0(\omega\tau) + i\frac{3\sigma_0}{2}\left[\frac{1-\cos(\omega\tau)\mathfrak{J}_0(\omega\tau)}{\omega\tau} - c\omega\tau\right], \qquad (12)$$

$$\sigma_\omega^{SD} = \frac{3\sigma_0}{2}\cos(\omega\tau)\mathfrak{J}_0(\omega\tau) + i\frac{3\sigma_0}{2}\left[\sin(\omega\tau)\mathfrak{J}_0(\omega\tau) - c\omega\tau\right]. \qquad (13)$$

Here $\sigma_0 = J_0/V_0$ is the dc conductance, J_0 is the dc current, $c \equiv 2C/3\sigma_0\tau$, C is the lateral *pin* junction capacitance, $\mathfrak{J}_0(\xi)$ is the Bessel function, and $\tau = l/v_W$ is the characteristic transit time of electrons and holes across the *i* section of length $2l$. Expressions (12) and (13) correspond to the cases of uniform and strongly nonuniform tunneling generation of the electron-hole pairs, respectively. In the former case, the electric field is constant, $E = V/2l$. In the latter case, it is assumed that the electric field is concentrated near the edges of the quasi-neutral *p* and *n* sections.

Figure 7 shows the real parts of the ac conductance σ_ω^{SD} as functions of the signal frequency $f = \omega/2\pi$ calculated for G-TUNNETTs with both uniform and strongly nonuniform electric-field distributions for different lengths of the *i* section. The phase diagrams, that is the pertinent $\mathrm{Im}\{\sigma_\omega^{SD}\}$ *vs.* $\mathrm{Re}\{\sigma_\omega^{SD}\}$ dependences for the device with a uniform electric-field distribution are also shown in Fig. 7. The real part exhibits a pronounced oscillatory behavior with frequency, including changes in sign. The frequency dependence of the imaginary part corresponds to the dominance of the inductive component at low frequencies (associated with the contribution of electrons and holes) and the capacitive component at high frequencies (due to the contribution of the geometrical capacitance). The relative values of the real part of the ac conductance at the first minima at $2l = 0.5$ μm is as follows: $\mathrm{Re}\{\sigma_\omega^{SD}\}/\sigma_0 \approx -0.034$ (at $f_1 = 1.75$ THz) and $\mathrm{Re}\{\sigma_\omega^{SD}\}/\sigma_0 \approx -0.141$ (at $f_1 = 1.25$ THz) for the uniform and nonuniform distributions, respectively. At $V_0 = 0.1$–1 V, for the above data one obtains $\mathrm{Re}\{\sigma_\omega^{SD}\} \approx -(0.71\text{–}2.26)\times10^{11}\ \mathrm{s}^{-1}$ and $\mathrm{Re}\{\sigma_\omega^{SD}\} \approx -(2.94\text{–}9.37)\times10^{11}\ \mathrm{s}^{-1}$. The

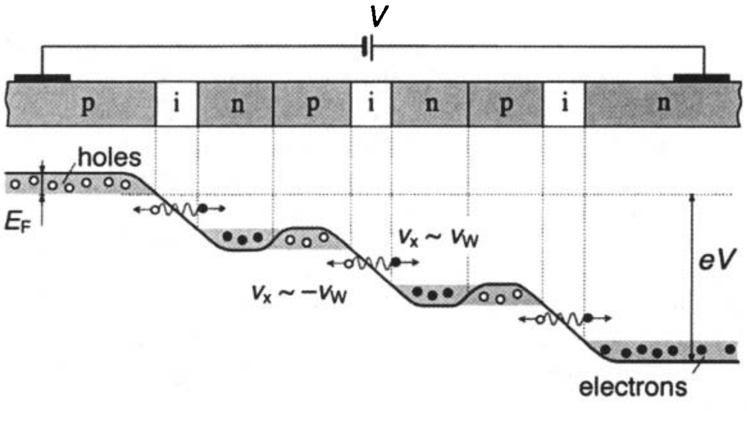


Figure 8. Schematic view of a cascaded G-TUNNETT structure and band diagram.

largest of these values, corresponding to several tens of mS/mm, markedly exceed those in some new concept THz devices based on different materials. The above estimates and Fig. 7 show that the real part of the ac conductance can be negative at the frequencies $f > 1$ THz in the G-TUNNETT structures with the *i* section length only moderately smaller than one micron. The G-TUNNETT can be used in THz oscillators with a complementary resonant cavity. Apart from the G-TUNNETTs with a single *pin* junction, structures with multiple cascaded *pin* junctions (see Fig. 8) exhibiting higher injection efficiency and better coupling with the electromagnetic radiation can be utilized. In the device structures based on the electrically induced *pin* junctions, such as injection lasers and G-TUNNETTs, the highly conducting split gates can serve as the waveguides.

4. Conclusions

We have considered new THz device concepts based on GLs, utilizing the specific features of the energy spectrum of graphene, namely, the zero energy gap and the constant velocity of electrons and holes. We demonstrated that novel THz sources can generate the electromagnetic radiation with the frequencies starting from about 1 THz at room temperatures.

Apart from the devices considered above, GLs can be used in THz photodetectors (see, for instance, Ref. 18). The controlled opening of the energy gap in graphene nanoribbons and graphene bilayers opens up prospects of high-performance THz transistors[19,20] and tunable phototransistors for THz radiation.[21,22]

References

1. C. Berger, Z. Song, T. Li, *et al.*, "Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics," *J. Phys. Chem.* **108**, 19912 (2004).
2. K. S. Novoselov, A. K. Geim, S. V. Morozov, *et al.*, "Two-dimensional gas of massless Dirac fermions in graphene," *Nature* **438**, 197 (2005).
3. A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, "The electronic properties of graphene," *Rev. Mod. Phys.* **81**, 109 (2009).
4. V. Ryzhii, M. Ryzhii, and T. Otsuji, "Negative dynamic conductivity of graphene with optical pumping," *J. Appl. Phys.* **101**, 083114 (2007).
5. F. Rana, "Graphene terahertz oscillators," *IEEE Trans. Nanotechnol.* **7**, 91 (2008).
6. A. Satou, F. T. Vasko, and V. Ryzhii, "Nonequilibrium carriers in intrinsic graphene under interband photoexcitation," *Phys. Rev. B* **78**, 115431 (2008).
7. A. A. Dubinov, V. Ya. Aleshkin, M. Ryzhii, T. Otsuji, and V. Ryzhii, "Terahertz laser with optically pumped graphene layers and Fabry-Perot resonator," *Appl. Phys. Express* **2**, 092301 (2009).

8. Yu. Miyamoto, H. Handa, E. Saito, *et al.*, "Raman-scattering spectroscopy of epitaxial graphene formed on SiC film on Si substrate," *e-J. Surf. Sci. Nanotech.* 7, 107 (2009).
9. P. Neugebauer, M. Orlita, C. Faugeras, A.-L. Barra, and M. Potemski, "How perfect can graphene be?" *Phys. Rev. Lett.* **103**, 136403 (2009).
10. E. D. Palik, *Handbook of Optical Constants of Solids*, New York: Academic Press, 1998, p. 192.
11. J. H. Strait, P. A. George, J. M. Dawlaty, S. Shivaraman, M. Chandrashekhar, F. Rana, and M. G. Spencer, "Emission of terahertz radiation from SiC," *Appl. Phys. Lett.* **95**, 051912 (2009).
12 F. Rana, P. A. George, J. H. Strait, J. Dawlaty, S. Shivaraman, M. Chandrashekhar, and M. G. Spencer, "Carrier recombination and generation rates for intravalley and intervalley phonon scattering in graphene," *Phys. Rev. B* **79**, 115447 (2009).
13. D. Sun, Z.-K. Wu, C. Divin, *et al.*, "Ultrafast relaxation of excited Dirac fermions in epitaxial graphene using optical differential transmission spectroscopy," *Phys. Rev. Lett.* **101**, 157402 (2008).
14. J. M. Dawlaty, S. Shivaraman, M. Chandrashekhar, F. Rana, and M. G. Spencer, "Measurement of ultrafast carrier dynamics in epitaxial graphene," *Appl. Phys. Lett.* **92**, 042116 (2008).
15. V. V. Cheianov and V. I. Fal'ko, "Selective transmission of Dirac electro ns and ballistic magnetoresistance of n-p junctions in graphene," *Phys. Rev. B* **74**, 041403 (2006).
16. M. Ryzhii and V. Ryzhii, "Injection and population inversion in electrically induced p-n junction in graphene with split gates," *Japan. J. Appl. Phys.* **46**, L151 (2007).
17. V. Ryzhii, M. Ryzhii, V. Mitin, and M. S. Shur, "Graphene tunneling transit-time terahertz oscillator based on electrically induced p-i-n junction," *Appl. Phys. Express* **2**, 034503 (2009).
18. F. T. Vasko and V. Ryzhii, "Photoconductivity of intrinsic graphene," *Phys. Rev. B* **77**, 195433 (2008).
19. M. Ryzhii, A. Satou, V. Ryzhii, and T. Otsuji, "High-frequency properties of a graphene nanoribbon field-effect transistor," *J. Appl. Phys.* **104**, 114505 (2008).
20. V. Ryzhii, M. Ryzhii, A. Satou, T. Otsuji, and N. Kirova, "Device model for graphene bilayer field-effect transistor," *J. Appl. Phys.* **105**, 104510 (2009).
21. V. Ryzhii, V. Mitin, M. Ryzhii, N. Ryabova, and T. Otsuji, "Device model for graphene nanoribbon phototransistor," *Appl. Phys. Express* **1**, 06302 (2008).
22. V. Ryzhii and M. Ryzhii, "Graphene bilayer field-effect phototransistor for terahertz and infrared detection," *Phys. Rev. B* **79**, 245311 (2009).

Directed Self-Assembly – A Controllable Route to Optical and Electronic Devices Based on Single Nanostructures

R. L. Williams, D. Dalacu, M. E. Reimer, K. Mnaymneh, V. Sazonova,
P. J. Poole, G. C. Aers, R. Cheriton, S. Frédérick, D. Kim,
J. Lapointe, P. Hawrylak, and M. Korkusiński
Institute for Microstructural Sciences, National Research Council
Ottawa, Canada, K1A 0R6

1. Introduction

Self-assembled semiconductor quantum dots offer an extremely attractive route to electronic structure control. Size, shape, composition and strain can all be used to tune the electronic structure of individual dots. Although useful, this versatility can also be a major drawback; particularly in devices where one would wish to couple to an individual dot or in situations where one would like many identical copies of a particular single dot structure. By the very nature of the self-assembled growth process, the characteristics of individual dots can vary widely and their spatial location is generally uncontrolled. In this chapter, we show how these limitations may be overcome and present examples in which control structures, such as optical microcavities or electrostatic gates, are constructed around individual quantum dots in an effort to determine their coupling to the optical field or to tune their electronic structure. Using such techniques one can engineer the symmetries of individual dots, introduce optical transitions that were previously forbidden, and so facilitate the construction of devices for the emission of single and entangled photon pairs.

2. Directed self-assembly

In certain compressively strained semiconductor materials systems, three-dimensional islands, or quantum dots (QDs), can be nucleated spontaneously in a so-called Stranski–Krastanow self-assembly process.[1] It is possible to gain considerable control over this nucleation process if one can influence the surface migration behavior of adatoms deposited upon the crystal surface during growth.[2-4] Figure 1 illustrates one possible route to achieve the necessary control over surface diffusion processes: crystal growth, in this case chemical beam epitaxy (CBE) of InP,[5] is initiated on an (001) oriented InP wafer that is patterned with openings in an otherwise continuous SiO_2 mask. Under the appropriate conditions, InP growth

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

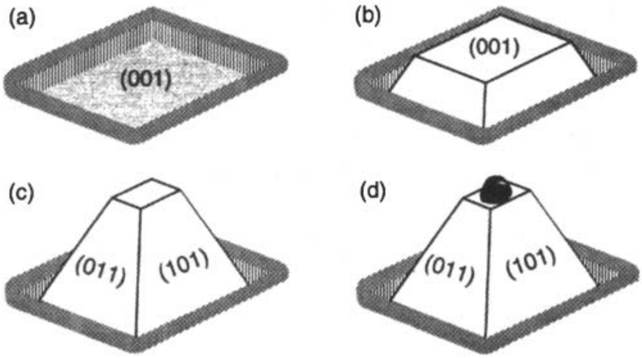


Figure 1. Schematic illustration of single, site-selected InAs/InP QD growth. In (a), the underlying InP substrate is patterned with square openings in an otherwise continuous SiO_2 masking layer. For openings with <100> edges, subsequent growth (b) produces square-based, pyramidal InP nanotemplates with {011} side facets. In (c), further InP growth results in a controlled reduction of the available (001) top surface area. In (d), deposition of InAs on the nanotemplate produces a single QD nucleated at the pyramid apex. Subsequent capping of the dot with InP produces a high-quality, optically-active InAs/InP QD whose position is known with high precision.

on such a masked substrate produces deposition only on the areas of exposed semiconductor substrate, resulting in defect free InP nanotemplates with a (001) top surface and sloping sidewalls.

The crystallographic orientation of these sidewalls is determined by the alignment of the oxide window: for example, square openings with edges directed along the <100> crystal directions will produce square based pyramidal InP nanotemplates with sloping {011} sidewalls. As the quantity of deposited InP increases, the nanotemplate height also increases, whilst the area of the (001) top surface decreases. In this manner, templates whose original dimension is of the order of a few square microns can be controllably grown until their top dimension approaches 50x50 nm; or the order of a typical QD dimension. If, at this point, the nanotemplate is exposed to the precursors for InAs growth, one finds that the deposited material migrates away from the {011} sidewalls to produce QD nucleation on the nanotemplate apex, as illustrated in Fig. 1. The beauty of this technique stems from the fact that the InP nanotemplate controlling quantum dot nucleation is prepared *in-situ*, in an ultrahigh-vacuum environment, with a size and geometry that is determined by the crystallography of the zincblende materials and by the quantity of InP deposited during nanotemplate growth. If the size and orientation of the original opening in the oxide is well controlled then the properties of the individual dots that are nucleated should also be well controlled. Figure 2 below illustrates the level of control that can be achieved with the nano-template, directed self-assembly technique.

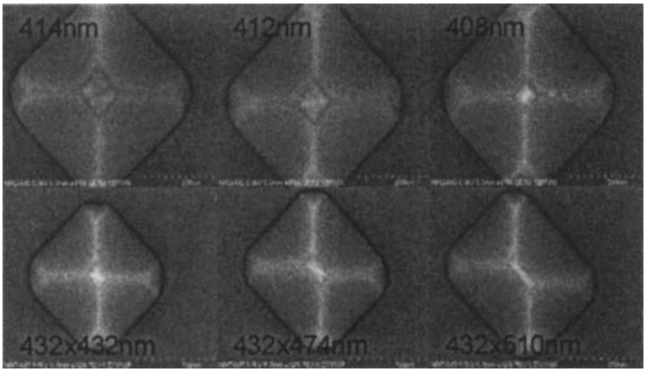


Figure 2. Scanning electron micrographs of uncapped InP nanotemplates with single InAs/InP QDs nucleated at the apex. Top row, left to right, shows the effect on QD lateral dimension of decreasing the area of the available (001) top surface. Single QDs tend to fill the available (001) area, allowing direct size control. Bottom row, left to right, shows the effect on quantum dot symmetry of elongating the InP nanotemplate. This control of quantum dot geometry allows one to engineer the symmetries of electronic states within the quantum dot.

Using the nanotemplate, directed self-assembly technique, not only can one control the spatial location of the dot, but by carefully designing the size and shape of the InP nanotemplate, one is able to control the size and shape of the nucleated QD. Control of the dot dimensions allows tuning of the electronic structure and thus the emission energies and energetic degeneracies. This is of particular relevance to entangled photon pair generation for example,[6-9] where the dot size and shape will determine whether the photon polarizations are entangled or merely correlated.[10]

3. Spectroscopic signatures

Single InAs/InP quantum dots can be designed for optical emission in the wavelength range between 1300 nm and 1550 nm (950 meV to 800 meV) and are consequently of interested for fiber-based optical communications.[11] Figure 3 shows typical luminescence[12] from a single InAs/InP quantum dot, as a function of increasing power from a HeNe pump laser.

At the lowest pump power (10 nW), a single optical emission line is observed corresponding to the annihilation of a single electron and single hole occupying the doubly degenerate s-shell of the quantum dot. It is this ground state emission that is of interest for single photon sources at telecom wavelengths and, as we shall see below, is directly tunable through choice of nanotemplate dimension. As a function of increasing pump power, successive electron-hole pairs are added to the quantum dot, revealing the rich many-body physics in play within these structures.[13]

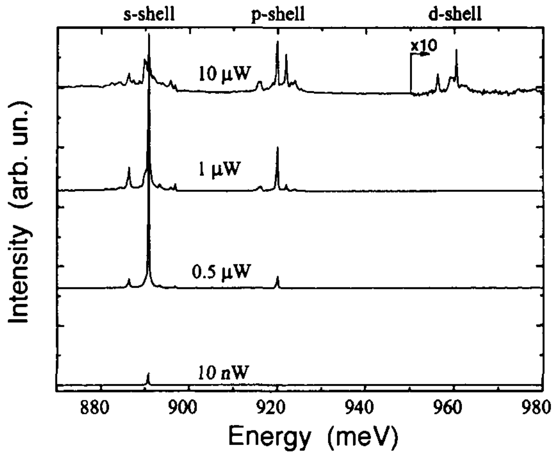


Figure 3. Luminescence from a single InAs/InP QD showing the development of electronic shell structure as a function of increasing optical pump power.

Single dot spectroscopy is facilitated by the above-described nanotemplate self-assembly process in that individual pyramidal templates can be spaced many hundreds of microns apart upon a substrate surface and easily imaged with an optical microscope. Luminescence from a single nanotemplate can then be excited and collected using the same microscope objective, before being spectrally dispersed within a grating spectrometer and detected using a multi-element detector. Figure 4 shows luminescence collected in this manner from a series of pyramidal templates with increasing base dimension. As the nanotemplate base dimension increases, the ground state luminescence energy from the quantum dot decreases as a result of diminishing electron and hole confinement.

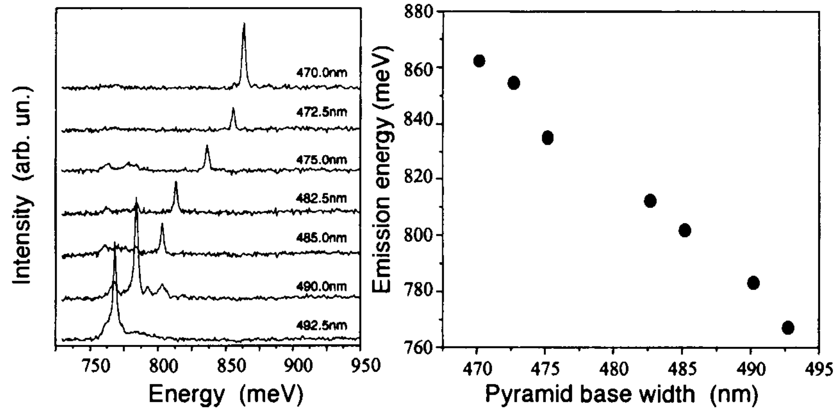


Figure 4. Ground state luminescence from single InAs/InP QDs on pyramidal nanotemplates of varying base dimension. The luminescence shifts to lower energy as a function of increasing base width, at a rate of approximately 4.5 meV/nm.

4. Application of electric fields

- *Vertical fields*

The quiescent charge state of a single QD can be controlled by the addition of Schottky gates to the sidewalls of a pyramidal nanotemplate, as illustrated in Fig. 5(a). In combination with an Ohmic back contact, Fig. 5(b), the four top gates can be used to apply an electric field in the vertical direction, raising or lowering the QD energy states with respect to the Fermi level.[14,15] At large negative gate bias ($V_G = -1.8$ V in Fig. 5(c)), the Fermi level lies below the QD electronic s-shell, producing an empty dot. In such circumstances, ground state optical emission from the dot corresponds to the neutral exciton X, as shown in the lowest spectrum of Fig. 5(c). At a gate bias of $V_G = -1.1$ V, the Fermi level lies within the electronic s-shell, resulting in single electron charging of the quantum dot and the characteristic signature of a trion X^-, in emission. When studied at high resolution, the optical emission from X and X^- can be differentiated by virtue of the small (10–100 μeV), exchange splitting present in the neutral exciton emission, producing a polarization-resolvable doublet. This splitting is absent for X^-, since the initial state in the transition corresponds to a spin singlet pair of electrons and the final state to an unpaired electron. With a further reduction in the negative gate bias, $V_G = 0$ in Fig. 5(c), an additional electron is added to the quantum dot to produce X^{2-}.

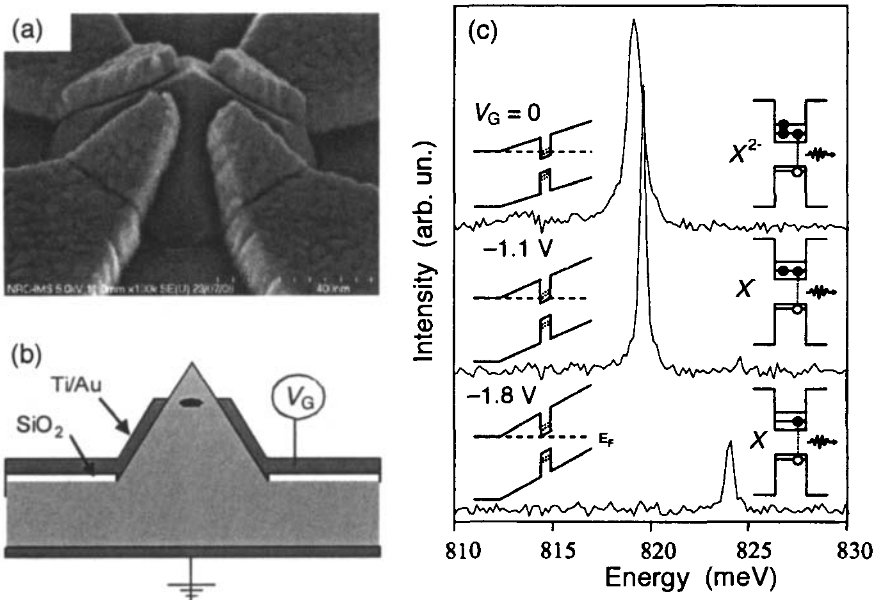


Figure 5. (a) Scanning electron micrograph of a pyramidal InP nanotemplate with four Schottky gates; (b) schematic illustration showing the top gate metal, SiO_2 dielectric, back gate and quantum dot position; (c) luminescence signal corresponding to three charge states of the quantum dot.

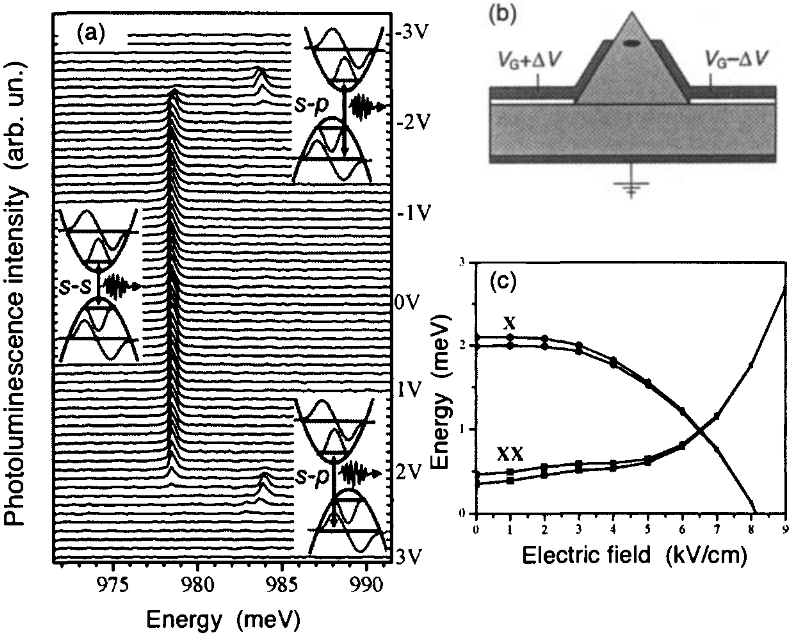


Figure 6. Effect of a lateral electric field on exciton and biexciton transitions within an InAs/InP QD. Once the vertical field has been used to select the desired charge state of a dot, the lateral field can be used to introduce new transitions, or to engineer degeneracies between existing ones.

Ground state luminescence from X^{2-} is again revealed as a doublet, corresponding to the two spin configurations of the final state (electrons in the s- and p-shells of the QD), although the low energy peak in Fig. 5 is very weak around 813 meV.

- *Lateral fields*

When differentially biased with respect to the back gate, the Schottky contacts of Fig. 6(b) can be used to apply a lateral electric field in the plane of the QD. Photoluminescence data from an InAs/InP QD in such a lateral field is shown in Fig. 6(a), along with insets showing the effects of this field upon the electron and hole in-plane, parabolic confining potential.[16,17,18] The lateral field separates the centers of electron and hole motion, introducing new optically allowed transitions and producing a red (Stark) shift of the single particle transition energies. An example of a normally forbidden optical transition can be observed around 984 meV in Fig. 6(a). This transition, which is observed under both directions of applied field, is associated with recombination from an s-shell electron and p-shell hole, a transition that is symmetry forbidden under zero applied bias.

Ideally, only ground state neutral exciton emission is observed under weak excitation conditions at zero applied bias. This is the situation in Fig. 6(a), with very weakly shifting neutral exciton emission observed around 978.5 meV. It is of interest to note that a single particle calculation of the expected Stark shift in these circumstances is grossly in error, with the predicted shift some two orders of magnitude larger than that observed in experiment. This phenomenon is explained by detailed many-body configuration interaction calculations, when it is observed that the reduction in electron-hole Coulomb interaction induced by charge separation almost completely compensates the expected single particle Stark shift for parabolic quantum dots. At first glance this compensation may be viewed as a detriment, since it removes one's ability to tune the exciton emission energy; a feature that may be of interest in resonant optical devices. However, if the mechanism that compensates the excitonic stark shift is applied to the biexciton transition, *XX*, one finds that this transition blue shifts with increasing lateral electric field, as shown in Fig. 6(c). At some electric field the *X* and *XX* transition energies are expected to become degenerate. Under these circumstances, where the binding energy of the biexciton has been removed, it is possible to identify two indistinguishable paths for photon pair emission from the biexciton to exciton to vacuum cascade, and photon entanglement that is of interest for quantum key distribution can be engineered.

5. Optical cavities

The CBE growth technique described above for the preparation of site-selected InAs/InP quantum dots is also amenable to subsequent growth steps that allow single pyramidal templates to be planarized and further processed to produce photonic crystal microcavity membranes. A schematic view of the structure is given in Fig. 7(a), along with an SEM image of the final cavity design in Fig. 7(b).

Single missing hole, InP membrane cavities of the type shown in Fig. 7(b) have been fabricated with quality factors around 30,000 and mode volumes around half a cubic wavelength.[19] With the ability to spatially align single InAs/InP QDs to the high field regions of these cavities, the potential exists to manufacture highly efficient single photon sources at telecom wavelengths, if one can match the cavity mode energy to the neutral exciton emission energy of the dot. At the present time, cavities with aligned single dots can be produced with quality factors of approximately 2,000, as shown in Fig. 7(c). The process used to planarize the pyramidal template produces a small local thickening of the final membrane around the template and shifts the cavity mode to slightly lower energy. This energy shift can be corrected and the cavity and exciton energies matched using a two stage process; for gross tuning, the surface of the cavity can be repeatedly oxidized and the resulting layer stripped, which has the effect of increasing the hole diameter and shifting the mode to higher energy.[20] For fine tuning the cavity mode can be dynamically tuned by condensing small amounts of nitrogen gas onto the cavity surface at low temperature. An example of such tuning is shown in Fig. 7(d).

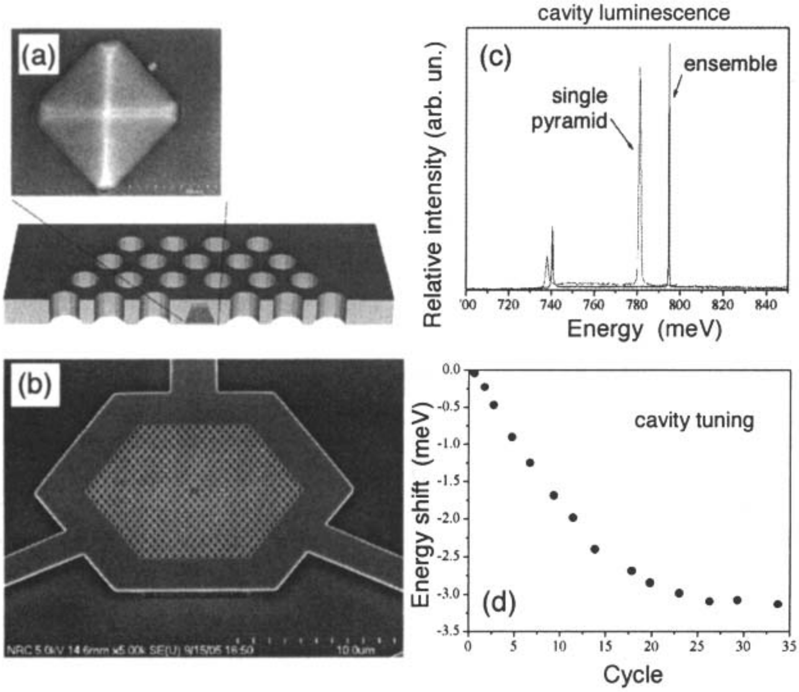


Figure 7. (a) Schematic illustration showing alignment of a site-selected single InAs/InP QD within a photonic crystal microcavity membrane; (b) scanning electron micrograph of fabricated cavity; (c) optical emission from a microcavity with either a QD ensemble or a spatially aligned single pyramidal template; (d) fine tuning of the cavity mode using successive cycles of N_2 condensation.

5. Conclusions

We discuss the application of directed self-assembly techniques to the production of single, site-selected InAs/InP QDs. Single dots can be produced that emit efficiently in the wavelength range suitable for fiber-based communications and control structures can be built around them, such as electrostatic gates and photonic crystal microcavities, that allow fine tuning of their electronic properties and of their coupling to the optical field. The techniques that we describe form the basis for a manufacturable technology for the production of non-classical light sources.

Acknowledgments

The authors would like to acknowledge the financial support of the Canadian Institute for Advanced Research, the Canadian Institute for Photonic Innovations, QuantumWorks, the Natural Sciences and Engineering Research Council, and the Business Development Bank of Canada.

References

1. I. N. Stranski and L. von Krastanow, *Akad. Wiss. Lit. Mainz Math.-Naturwiss. Kl. IIb* **146**, 797 (1939).
2. R. Williams, G. Aers, P. Poole, J. Lefebvre, D. Chitrani, and B. Lamontagne, "Controlling the self-assembly of InAs/InP quantum dots," *J. Crystal Growth* **223**, 321 (2001).
3. J. Lefebvre, P. J. Poole, J. Fraser, G. C. Aers, D. Chithrani, and R. L. Williams, "Self-assembled InAs quantum dots on InP nano-templates," *J. Crystal Growth* **234**, 391 (2002).
4. P. J. Poole, G. C. Aers, A. Kam, D. Dalacu, S. Studenikin, and R. L. Williams, "Selective growth of InP/InGaAs <010> ridges: Physical and optical characterization," *J. Crystal Growth* **310**, 1069 (2008).
5. P. J. Poole, J. McCaffrey, R. L. Williams, J. Lefebvre, and D. Chitrani, "Chemical beam epitaxy growth of self-assembled In/InP quantum dots," *J. Vac. Sci. Technol. B* **19**, 1467 (2001).
6. R. Seguin, A. Schliwa, S. Rodt, K. Pötchke, U. W. Pohl, and D. Bimberg, "Size-dependent fine-structure splitting in self-organized InAs/GaAs quantum dots," *Phys. Rev. Lett.* **95**, 257402 (2005).
7. R. J. Young, R. M. Stevenson, A. J. Shields, *et al.*, "Inversion of exciton level splitting in quantum dots," *Phys. Rev. B* **72**, 113305 (2005).
8. A. Greilich, M. Schwap, T. Berstermann, *et al.*, "Tailored quantum dots for entangled photon pair creation," *Phys. Rev. B* **73**, 045323 (2006).
9. N. Chauvin, B. Salem, G. Bremond, G. Guillot, C. Bru-Chevallier, and M. Gendry, "Size and shape effects on excitons and biexcitons in single InAs/InP quantum dots," *J. Appl. Phys.* **100**, 073702 (2006).
10. C. Santori, D. Fattal, M. Pelton, G. S. Solomon, and Y. Yamamoto, "Polarization-correlated photon pairs from a single quantum dot," *Phys. Rev. B* **66**, 045308 (2002).
11. D. Chithrani, R. L. Williams, J. Lefebvre, P. J. Poole, and G. C. Aers, "Optical spectroscopy of single, site-selected InAs/InP self-assembled quantum dots," *Appl. Phys. Lett.* **84** 978 (2004).
12. D. Kim, H. Lefebvre, J. Lapointe, M. E. Reimer, J. McKee, P. J. Poole and R. L. Williams, "Optical spectroscopy of single, planar, self-assembled InAs/InP quantum dots," *phys. stat. sol. (c)* **3**, 3840 (2006).
13. W. Sheng and P. Hawrylak, "Atomistic theory of electronic and optical properties of InAs/InP self-assembled quantum dots on patterned substrates," *Phys. Rev. B* **72**, 035326 (2005).
14. M. E. Reimer, D. Dalacu, J. Lapointe, *et al.*, "Single electron charging in deterministically positioned InAs/InP quantum dots," *Appl. Phys. Lett.* **94**, 011108 (2009).
15. M. E. Reimer, W. R. McKinnon, J. Lapointe, *et al.*, "Towards scalable gated quantum dots for quantum information applications," *Physica E* **40**, 1790 (2008).
16. M. E. Reimer, M. Korkusinski, D. Dalacu, *et al.*, "Prepositioned single quantum dot in a lateral electric field," *Phys. Rev. B* **78**, 195301 (2008).

17. M. E. Reimer, M. Korkusinski, D. Dalacu, *et al.*, "Voltage induced hidden symmetry and photon entanglement generation in a single, site-selected quantum dot," unpublished (arXiv:0706.1075v1 [quant-ph]).
18. M. Korkusinski, M. E. Reimer, R. L. Williams, and P. Hawrylak, "Engineering photon cascades from multiexciton complexes in a self-assembled quantum dot by a lateral electric field," *Phys. Rev. B* **79**, 035309 (2009).
19. S. Frédérick, D. Dalacu, J. Lapointe, P. J. Poole, G. C. Aers, and R. L. Williams, "Experimental demonstration of high quality factor, x-dipole modes in InAs/InP quantum dot photonic crystal microcavity membranes," *Appl. Phys. Lett.* **89**, 091115 (2006).
20. D. Dalacu, S. Frédérick, P. J. Poole, G. C. Aers, and R. L. Williams, "Postfabrication fine-tuning of photonic microcavities in InAs/InP quantum dot membranes," *Appl. Phys. Lett.* **87**, 151107 (2005).

Part IV

Sensors, Detectors, and Energy

This page intentionally left blank

4 Sensors, Detectors, and Energy

Continuing the central theme of focusing on future trends, the participants of this gathering and the contributors to this part of the book ventured beyond the current main path of the IT evolution – information processing and transmission – to the acquisition of information via sensing.

An inspiring Canadian effort on sensing at a very large scale – office buildings, hospitals, airports, shopping plazas, and over a full range of pollutants and gases – was presented and generated much enthusiasm. Another new topic in sensing was covered in a mini-symposium on recently-developed high-energy particle detectors. At the other end of the energy spectrum, we have continued the pursuit of very low energy infrared detection, this time featuring a proposed bolometric technology based on the VO_2 phase transition.

A strong advocacy was again made for large-area distributed electronics – macroelectronics, as well as for new energy sources.

Perhaps, in more ways than one, the emphasis on sensing signal the beginning of a paradigm shift in sensing and detection, long-advocated at the FTM meetings: real-time adaptive collection and read-out of spectral, spatial, directional, and substance-differentiated information.

Contributors

4.1	Zhong He
4.2	Serge Luryi and Arsen Subashiev
4.3	Arsen Subashiev and Serge Luryi
4.4	M. A. Alam, N. Pimparkar, and B. Ray
4.5	M. D'Iorio
4.6	A. Sergeev, V. Mitin, L. H. Chien, and N. Vagidov
4.7	M. Gurvitch, S. Luryi, A. Polyakov, and A. Shabalov
4.8	Bo-Kuai Lai, A. C. Johnson, H. Xiong, C. Ko, and S. Ramanathan

This page intentionally left blank

Three-Dimensional Position-Sensitive Wide Bandgap Semiconductor Gamma-Ray Imaging Detectors

Zhong He
Nuclear Engineering and Radiological Sciences Department
The University of Michigan, Ann Arbor, MI 48109-2104, U.S.A.

1. Introduction

Among the three major types of gamma-ray detectors – based on scintillators, gases or semiconductors – semiconductor detectors can offer the best energy resolution as well as the high detection efficiency of a solid state detector. High-purity germanium (HPGe) detectors have been the gold standard for high-resolution gamma-ray spectroscopy since the mid 1970's,[1] but they must be operated at cryogenic temperatures of around 77 K (about –196 °C) due to germanium's small bandgap of ~0.7 eV. The requirement of cryogenic cooling has made the deployment of HPGe detectors difficult for many applications. Although mechanically cooled HPGe detectors have become available in recent years, there has been strong interest in finding alternative detector technologies to minimize power consumption, instrument weight and risks associated with failures of cooling and vacuum systems of HPGe detectors.

Wide bandgap semiconductors with high atomic numbers and densities are desired for both efficient gamma-ray detection and the convenience of room-temperature operation. Several wide bandgap semiconductors, such as CdTe, CdZnTe, HgI_2 and TlBr have attracted the most attention. However, charge-carrier trapping within the detector material, in particular the trapping of holes, causes incomplete charge collection. Despite continued efforts since the 1960's to identify impurities that trap charges and improve crystal quality, the mean free drift length of holes remains less than 1 mm at practical bias voltages in wide bandgap semiconductors having higher atomic numbers. The severe trapping of holes makes the signal amplitude Q, recorded on the cathode or the anode of a conventional planar-electrode detector, proportional to the product of the number of electron-hole pairs n_0 generated by gamma-ray interaction and the normalized drift distance z of electrons relative to the total distance between the cathode and the anode. This relationship can be expressed as $Q \propto (n_0 e)\cdot z$ where e is the electron charge and $0 \leq z \leq 1$. If we do not know the depth of interaction z, the true gamma-ray energy deposition $E_0 \propto n_0 e$ cannot be determined from each measured signal amplitude Q. Figure 1 shows the Cs-137 energy spectrum recorded from the cathode electrode of one 2x2x1.5 cm^3 CdZnTe detector (#4E-1). The Cs-137 source emits x-rays between 32 and 38 keV, and monoenergetic gamma rays at 662 keV. A clear full-energy peak can be seen at ~35 keV (around 340 ADC

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

channels), but only a high-energy cut-off can be seen at 662 keV. This is because low-energy x-rays are absorbed within a thin layer of detector material, thus their depth of interaction is very close to the cathode surface when the source was located on the cathode side. In this case, all electrons travel through the entire thickness of the CdZnTe detector, thus the relative depth of interaction z is close to 1, leading to $Q \propto (n_0 e) \propto E_0$. In contrast, the depth of interaction for higher-energy 662 keV gamma rays can be anywhere between the cathode and the anode surfaces, thus $0 \leq z \leq 1$. This is why the recorded signal amplitude Q can vary continuously from ADC channel around 280, corresponding to $E_0 \sim 0$, up to about 1300 ADC channels corresponding to $E_0 = 662$ keV.

Many pulse-shaping techniques have been explored to correct this signal deficit,[2,3] but it is not possible to achieve good energy resolution and high detection efficiency simultaneously, especially on detectors with thickness of 1 cm or more.[4] Single-polarity charge-sensing techniques, in which the output signal depends on the drift and collection of charge carriers of only one polarity (electrons or holes), have shown promise to overcome this problem. Our group proposed the three-dimensional (3D) position-sensing technology in 1998[5] and subsequently demonstrated better than 1% FWHM energy resolution for 662 keV gamma rays on CdZnTe detectors with dimensions up to 2x2x1.5 cm^3.[6] Energy resolutions of about 1.6% FWHM have been demonstrated on HgI_2 detectors with dimensions of up to 1.8x1.8x1.5 cm^3.[7]

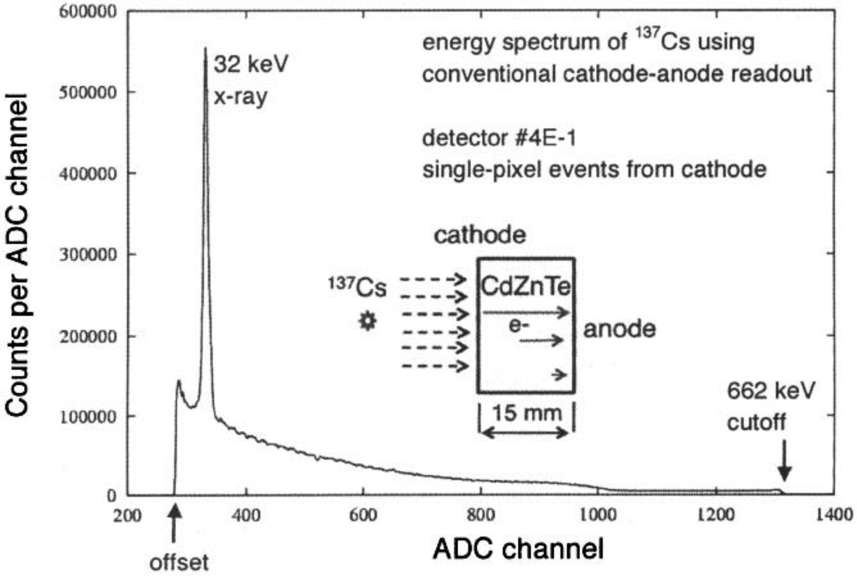
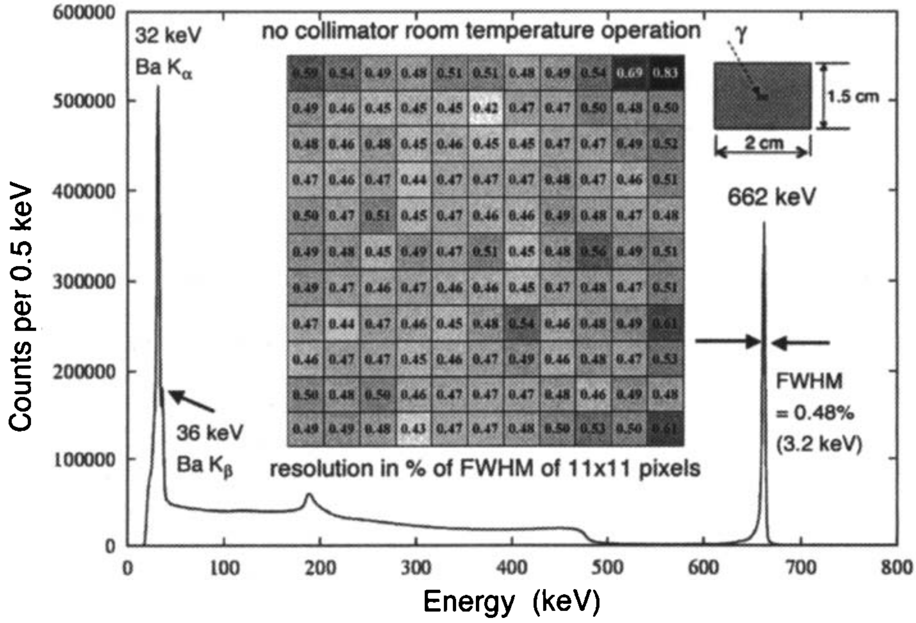


Figure 1. Measured Cs-137 energy spectrum obtained from the cathode electrode. The x-rays and gamma rays are incident on the cathode surface as illustrated in the figure. A photopeak is obtained at 32–38 keV, but only a continuous spectrum with a high-energy cutoff is recorded for the 662 keV gamma ray. The CdZnTe detector (#4E-1) has dimensions of 2x2x1.5 cm^3.

During the past ten years, the Department of Energy NA-22 office has nurtured 3D CdZnTe detector technology by maintaining steady support for CdZnTe material development, device fabrication and application-specific integrated circuitry (ASIC) development. Since 2006, the development of CdZnTe gamma-ray imaging detectors has moved to a new era. The DoD's Defense Threat Reduction Agency (DTRA) has significantly invested in developing large volume (up to 2x2x1.5 cm^3) CdZnTe detectors and ASIC-based low-noise high-performance electronic readout systems under the Polaris project at the University of Michigan. The goals of the Polaris project are to explore the best spectroscopic performance up to 3 MeV and develop real-time gamma-ray imaging capability with high detection efficiency. The Domestic Nuclear Detection Office (DNDO) of the Department of Homeland Security has developed 3D CdZnTe-based detector systems in the intelligent personal radiation locator (IPRL) project, which has focused on detector packaging and sensor-network capabilities. The Department of Energy NA-22 office has supported the GammaTracker project at Pacific Northwest National Laboratory to construct the first hand-held isotope identifier with efficiency comparable to commercial HPGe detectors for field operation. This is also based on 3D CdZnTe detector technology. These new developments have shown the great potential of large-volume wide bandgap semiconductor gamma-ray imaging detectors. This paper summarizes the latest technology development on 3D-position-sensitive wide bandgap semiconductor gamma-ray imaging spectrometers at the University of Michigan.

2. Gamma-ray spectroscopy

To overcome the problem of severe hole trapping, an array of 11x11 pixel anodes is fabricated on the anode surface to replace the conventional single plane anode electrode.[5] The signal amplitude is proportional to the number of electrons collected on each pixel anode by reading signals from individual pixel anodes using multi-channel low-noise charge-sensitive ASIC chip. In addition, the drift times of electrons can be obtained from the time intervals between the cathode's and pixel anodes' trigger times provided by the ASIC. The depth of interaction of each individual energy deposition can be determined from the drift time measurement. Knowledge of the depth of interaction enables the correction of the signal deficit due to the trapping of electrons, which, although small, is still significant to degrade energy resolution without depth correction. Figure 2 shows the latest spectroscopic performance measured on one 2x2x1.5 cm^3 CdZnTe detector (#4E-1) using the 3D-position-sensitive readout technology,[6] operated at ambient temperature in our laboratory. An energy resolution of 0.48% FWHM at 662 keV for single-pixel events was obtained from the entire detector volume. The energy resolution (FWHM in percent at 662 keV) from each anode pixel (after depth correction) is also shown in the same figure. Single-pixel events, in which all of the electrons generated by an incident gamma ray are collected by a single anode pixel, give the best energy resolution and account for about 30% of the 662 keV photopeak events. This energy resolution is approximately within a factor of two of the energy resolution achievable by high-purity Ge detectors, which is very close to the theoretical limit achievable by semiconductor detectors. Comparing to the result shown in Fig. 1, a clear full-energy peak is obtained at 662 keV on the same detector. This comparison demonstrates the advantage of the single-polarity charge-sensing technique over the conventional planar-electrode readout on detectors suffering charge trapping. If all 1-, 2-, 3-, and more than 3-pixel events are included, the overall energy resolution is 0.69% FWHM at 662 keV on this CdZnTe detector (#4E-1). This energy resolution includes the noise contribution of the ASIC, which is about 2 keV (0.3%) FWHM on each channel connected to anode pixels. The dynamic range of each ASIC channel is about 3 MeV. Gamma rays with higher energy than 3 MeV can be recorded by summing up energy depositions from multiple pixel anodes if the energy deposition underneath each pixel anode is smaller than 3 MeV. In addition, photopeaks at 32 and 36 keV, corresponding to K_α and K_β x-rays of barium in the Cs-137 source, are clearly separated in Fig. 2. The better energy resolution at low energies, compared to the result shown in Fig. 1, is due to the fact that the electronic noise of each pixel anode channel is about 2 keV FWHM, lower than that of the cathode channel which is about 6 keV FWHM.



0.59	0.54	0.49	0.48	0.51	0.51	0.48	0.49	0.54	0.69	0.83
0.49	0.46	0.45	0.45	0.45	0.42	0.47	0.47	0.50	0.48	0.50
0.48	0.46	0.48	0.45	0.46	0.45	0.45	0.47	0.47	0.49	0.52
0.47	0.46	0.47	0.44	0.47	0.47	0.47	0.48	0.47	0.46	0.51
0.50	0.47	0.51	0.45	0.47	0.46	0.46	0.49	0.48	0.47	0.48
0.49	0.48	0.45	0.49	0.47	0.51	0.45	0.48	0.56	0.49	0.51
0.49	0.47	0.46	0.47	0.46	0.46	0.45	0.47	0.48	0.47	0.51
0.47	0.44	0.47	0.46	0.45	0.48	0.54	0.46	0.48	0.49	0.61
0.46	0.47	0.47	0.45	0.46	0.47	0.49	0.46	0.48	0.47	0.53
0.50	0.48	0.50	0.46	0.47	0.47	0.47	0.48	0.46	0.49	0.48
0.49	0.49	0.48	0.43	0.47	0.47	0.48	0.50	0.53	0.50	0.61

Figure 2. Single-pixel energy spectrum of Cs-137, obtained using one BNL ASIC readout from all 121 anode pixels of a $2 \times 2 \times 1.5$ cm^3 CdZnTe detector (#4E-1) at room temperature. The energy resolution (FWHM in % at 662 keV) *vs.* the location of 11x11 anode pixels is also shown. The grey color scale is 0.4–1.0% FWHM.

Three-dimensional position-sensing technology has also been applied to HgI_2 detectors, an alternative wide bandgap semiconductor having higher stopping power (due to higher atomic number and density). Figure 3 shows the overall single-pixel Cs-137 energy spectrum obtained from an array of eighteen 1.8x1.8x1 cm^3 HgI_2 detectors.[7] Although this performance is not as good as that of the CdZnTe detector, significant progress has been made when compared to what was possible ten years ago – typically 2–3% FWHM energy resolution at 662 keV from detectors of 1 to 2 mm thick. This result also demonstrates that a large number of modular detectors, in this case an array of eighteen, can be assembled together and operated as a single instrument, providing the much higher detection efficiency needed for higher-energy gamma rays. During 2008–9, five 1.8x1.8x1.5 cm^3 HgI_2 detectors, manufactured by the Constellation Technology Corporation, have demonstrated overall single-pixel event energy resolution better than 2% FWHM at 662 keV. Close to 1% FWHM energy resolution at 662 keV was observed over a significant fraction of the detection volume on several 1.5-cm thick HgI_2 detectors, showing the potential of HgI_2 detectors to achieve much better spectroscopic performance with higher detection efficiency.[7]

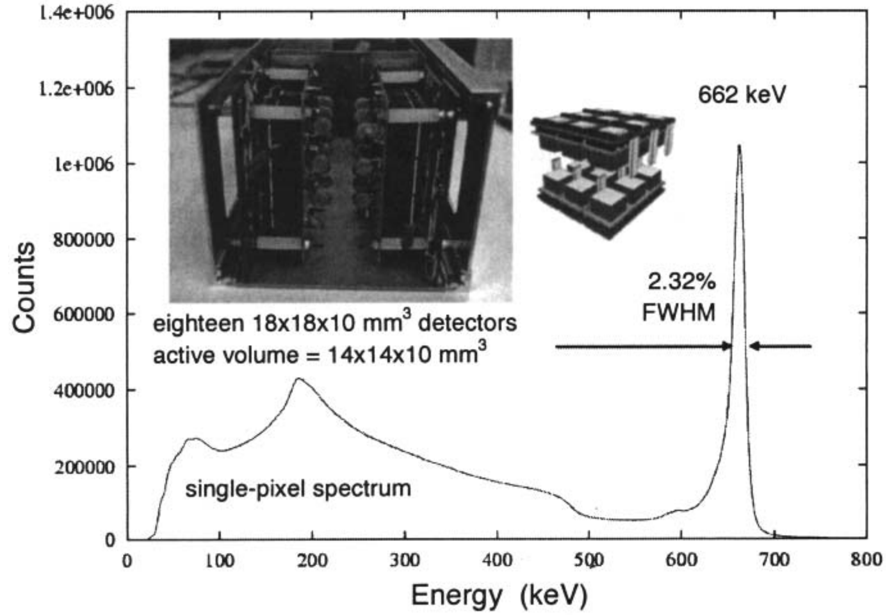


Figure 3. Single-pixel energy spectrum of Cs-137 obtained from an array of eighteen 1.8x1.8x1.0 cm^3 HgI_2 detectors, each connected to one ASIC manufactured by Gamma Medica Ideas. An array of 11x11 pixel anodes was fabricated over an area of 14x14 mm^2 centered on the 18x18 mm^2 anode surface.

3. Gamma-ray Compton imaging

Three-dimensional position sensing technology has enabled CdZnTe detectors to achieve energy resolution close to the theoretical limit on semiconductor detectors despite the problems of charge trapping; it also provides the energy deposition and position for individual radiation interactions. This energy-position information makes it possible to perform Compton gamma-ray imaging even on a single semiconductor detector without the use of any collimator[8] and to identify unique signatures and characteristics of different types of radiation interactions. As an example, gamma-ray polarization can be determined by measuring the angular preference of scattered gamma rays within a single 3D-position-sensitive CdZnTe detector.[9] Figure 4 shows one of our experimental results for Compton gamma-ray imaging in which three different gamma-ray sources (Ba-133, Cs-137 and Co-60) were placed at different locations in our laboratory. All of the sources were correctly imaged, identified and located. The gamma-ray images (corresponding to each isotope) appear over a 360° optical panorama image of our laboratory.[10]

We experimentally investigated the accuracy of locating a 6.6 μCi Cs-137 source at a distance of 1 meter from the center of a single 2x2x1.5 cm^3 3D CdZnTe detector. The count rate was about 5.5 counts per minute within the energy window of 600 to 720 keV. The reconstructed image pixel size was 5° due to a very coarse angular grid selected for a fast real-time imaging demonstration. Our experiment showed that a 5° angular standard deviation on the estimate of the source's angular position can easily be obtained with a small number of events (< 50), using simple back-projection image reconstruction or maximum likelihood position estimation techniques. The actual accuracy may easily be better than this because the limit of the image pixel size was reached.[11] It has been demonstrated that radiation source detection can be improved by combining the spectroscopic and imaging information.[12]

To demonstrate that a single 2x2x1.5 cm^3 3D CdZnTe detector can detect and image very weak radiation sources, the natural gamma-ray background in the range

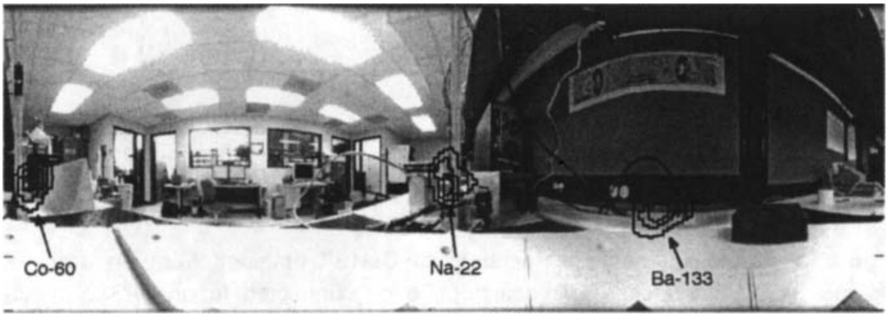


Figure 4. Three isotope sources (Ba-133, Na-22, and Co-60) were correctly located, identified and imaged simultaneously in our laboratory using a single 2x2x1.5 cm^3 CdZnTe detector.

of 550–650 keV was measured using a single 2x2x1.5 cm^3 CdZnTe detector (#4E-5) located inside a lead cave (see Fig. 5). Four of the six detector surfaces were shielded with lead blocks at least two inches thick, while the remaining two opposite sides were open. Figure 5 shows two hot regions in the reconstructed image using the simple back-projection technique in the 550–650 keV range, corresponding to the correct directions of the two openings.[10] All radiation sources were removed from the area so that the two hot gamma regions were from natural gamma-ray background emitted by building materials in our laboratory. Using a more advanced maximum-likelihood-based energy-imaging integrated deconvolution (EIID),[13] we can obtain the best estimate of the incident gamma-ray spectrum (before gamma rays strike the detector surface) from each direction within the entire 4π field of view. Figure 6 shows the EIID deconvolved energy spectra for a region with 30° half-angle centered at one of the two openings and from the (cathode) direction covered with lead shielding. The gamma-ray lines at 511 and 609 keV from Bi-214, and 1461 keV from K-40 in the concrete of the building can be clearly seen above the background noise in the direction of the opening. The heavy shielding produces a very low count rate, which resulted in

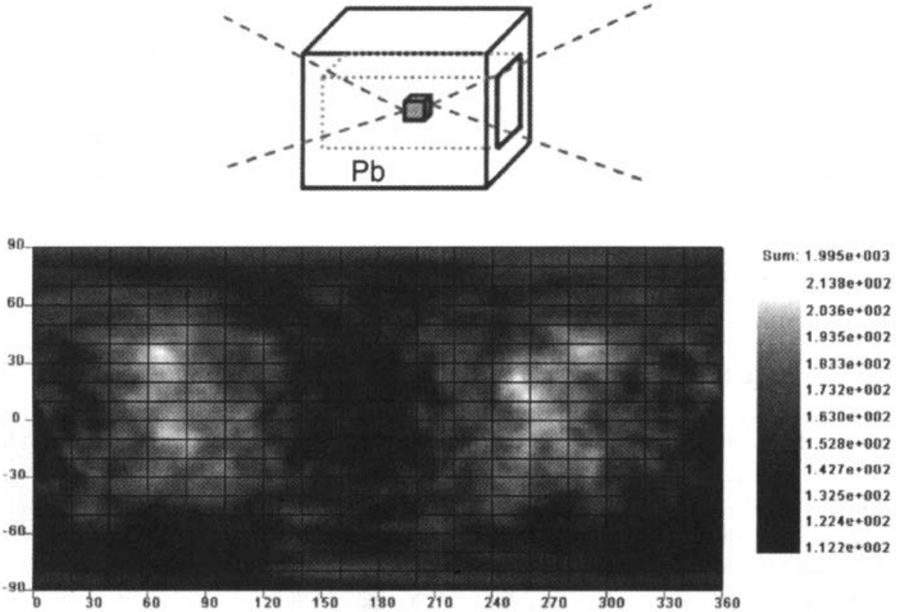


Figure 5. Top: illustration of a single 2x2x1.5 cm^3 CdZnTe detector located inside a lead cave with two openings opposite to each other. Bottom: the reconstructed image using simple back-projection in the 550–650 keV energy range showed that the distribution of natural gamma-ray background (no source present) can be correctly imaged. The horizontal axis is the polar angle and the vertical axis is the azimuthal angle.

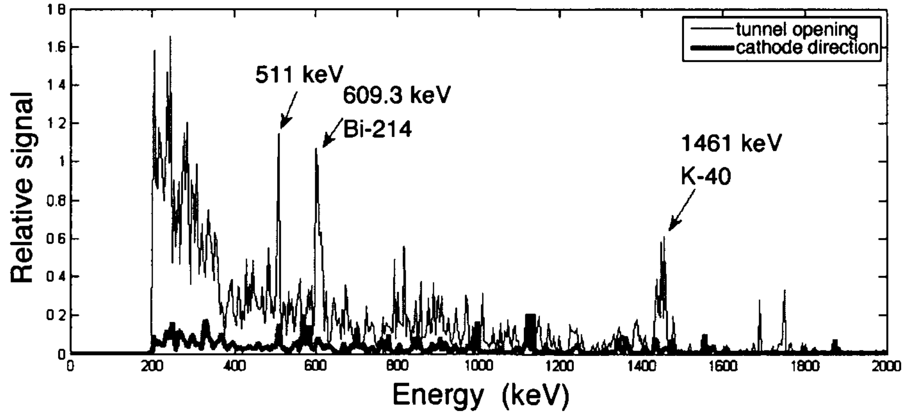


Figure 6. The deconvolved energy spectrum using energy-imaging integrated deconvolution (EIID) for a 30° half-angle region in the cave opening (thin line) and from the cathode direction which was shielded by lead blocks (thicker line). The vertical axis is proportional to recorded count.

these noisy spectra. This combination of gamma-ray spectroscopy and imaging capability will enable detection of nuclear isotope sources with higher sensitivity than spectroscopic information alone. For example, we have experimentally demonstrated that a 3D CdZnTe imaging detector can see a bottle of no-sodium salt (rich in K-40) by identifying a "hot" localized source in the reconstructed image at 1461 keV gamma-ray energy. Discrimination between a region with elevated background and a localized source of K-40 is impossible using only a gamma-ray spectrometer if the background radiation is unknown, since K-40 also exists in the natural background.

4. Conclusions

Three-dimensional-position-sensitive semiconductor gamma-ray imaging detectors have been developed over the past decade. The energy resolution has been improved from 1.7% FWHM at 662 keV for single-pixel events on one 1x1x1 cm^3 CdZnTe detector in 1998[5] to 0.48% FWHM on one 2x2x1.5 cm^3 CdZnTe detector in 2009,[6] a three-fold improvement on energy resolution and a six-fold increase in the detection volume. This technique has enabled the best CdZnTe detector to achieve energy resolution within a factor of two of the theoretically achievable spectroscopic performance of semiconductor detectors. Real-time gamma-ray Compton imaging has been demonstrated on 3D CdZnTe detectors. This gamma-ray imaging capability together with excellent spectroscopic performance enables higher sensitivity source detection, search and identification.

References

1. Glenn Knoll, *Radiation Detection and Measurement*, 3rd ed., New York: Wiley & Sons, 2001.
2. L. T. Jones and P. B. Woollam, "Resolution improvement in CdTe gamma detectors using pulse-shape discrimination, *Nucl. Instr. Meth.* **124**, 591 (1975).
3. M. Richter and P. Siffert, "High resolution gamma-ray spectroscopy with CdTe detector systems," *Nucl. Instr. Meth. Phys. Res. A* **322**, 529 (1992).
4. Zhong He, "Review of the Shockley-Ramo theorem and its application in semiconductor gamma-ray detectors," *Nucl. Instr. Meth. Phys. Res. A* **463**, 250 (2001).
5. Z. He, W. Li, G. F. Knoll, *et al.*, "3-D position-sensitive CdZnTe gamma-ray spectrometers," *Nucl. Instr. Meth. Phys. Res. A* **422**, 173 (1999).
6. Feng Zhang *et al.*, "Characterization of the H3D ASIC readout system and the 6 cm^3 3-D position sensitive CdZnTe detectors," to be submitted to *Nucl. Instr. Meth. Phys. Res. A* (2009).
7. Zhong He and Feng Zhang, "Gamma-ray spectroscopic performance of an array of HgI_2 detectors," to be submitted to *Nucl. Instr. Meth. Phys. Res. A* (2009).
8. D. Xu, Z. He, C. E. Lehner, and F. Zhang, "4π Compton imaging with single 3-D position sensitive CdZnTe detector," *Proc. SPIE* **5540**, 144 (2004).
9. D. Xu, Z. He and F. Zhang, "Detection of gamma ray polarization using a 3D position-sensitive CdZnTe detector," *IEEE Trans. Nucl. Sci.* **52**, 1160 (2005).
10. Weiyi Wang *et al.*, "Image reconstruction of the natural radiation background," to be submitted to *Nucl. Instr. Meth. Phys. Res. A* (2009).
11. C. G. Wahl and Z. He, "Sensitivity of gamma-ray source detection using 3D-position-sensitive semiconductor detectors," *Conf. Record IEEE Nucl. Sci. Symp.* (2008), pp. 3334–8.
12. C. G. Wahl and Z. He, "Point-source detection using energy and imaging information from 3D-position-sensitive semiconductor detectors," to appear in *Conf. Record IEEE Nucl. Sci. Symp.* (2009).
13. D. Xu and Z. He, "Gamma-ray energy-imaging integrated spectral deconvolution," *Nucl. Instr. Meth. Phys. Res. A* **574**, 98 (2007).

This page intentionally left blank

Semiconductor Scintillator for Three-Dimensional Array of Radiation Detectors

Serge Luryi and Arsen Subashiev
Dept. of Electrical and Computer Engineering, SUNY–Stony Brook
Stony Brook, NY 11794-2350, U.S.A.

1. Introduction

There are two large groups of solid-state radiation detectors that dominate the area of ionizing radiation measurements: scintillation detectors and semiconductor diodes. The scintillators detect high-energy radiation through generation of light, which is subsequently registered by a photodetector that converts light into an electrical signal. Semiconductor diodes employ reverse biased *pn* junctions where the absorbed radiation creates electrons and holes, which are separated by the junction field thereby producing a direct electrical response. Both groups of detectors are extensively reviewed in the classical treatise by Knoll.[1]

Most scintillators reported in the literature are implemented in wide-gap insulating materials doped ("activated") with radiation centers. A classic example of a solid-state scintillator is sodium iodide activated with thallium (NaI:Tl), introduced by Hofstadter[2] more than 60 years ago.

The typical band diagram of a solid-state scintillator is shown in Fig. 1. Because of the much longer wavelength of the scintillation associated with the activator energy levels – compared to the interband absorption threshold – the insulating scintillators are very transparent to their own luminescence. However, this advantage comes at a price in the transport of carriers to the activator site. Individual carriers have very poor mobility in insulators and transport efficiency

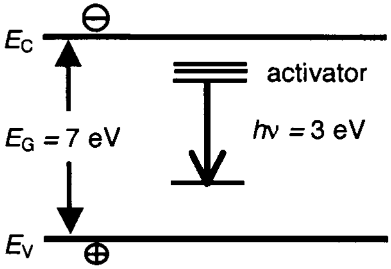


Figure 1. Conventional solid-state scintillator. Schematic energy band diagram of the thallium activated sodium iodide. Electrons and holes generated across the bandgap E_G migrate to the nearest activation site and recombine there, emitting much lower photon energy $h\nu$.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

requires that the generated electrons and holes form excitons and travel to the radiation site as neutral entities. Therein lies a problem. The energy resolution even in the best modern scintillators does not compare well with that in semiconductors.[4] One of the fundamental reasons for poor resolution is that the luminescent yield in dielectric scintillators is controlled by reactions that are nonlinear in the density of generated electron-hole pairs, such as the formation of excitons at low densities and the Auger recombination at high densities.[5-8]

It is important to stress that such nonlinear processes do not exist in direct-gap doped semiconductors, where interaction with gamma radiation induces *minority* carriers, while the concentration of majority carriers does not measurably change. Every reaction on the way to luminescence, including Auger recombination, is *linear* with respect to the concentration of minority carriers. One can therefore expect, as a matter of principle, that doped semiconductor scintillators will not exhibit effects of non-proportionality and their ultimate energy resolution could be on par with that of diode detectors implemented in the same material.

Normally, scintillators are not made of semiconductor material. The key issue in implementing a semiconductor scintillator is how to make the material transmit its own infrared luminescence, so that photons generated deep inside the semiconductor slab could reach its surface without tangible attenuation. However, semiconductors are usually opaque at wavelengths corresponding to their radiative emission spectrum. Our group has been working on the implementation of high-energy radiation detectors based on direct-gap semiconductor scintillator wafers, like InP or GaAs. For the exemplary case of InP, the scintillation spectrum is a band of wavelengths near 920 nm. The original idea was to make InP relatively transparent to this radiation by doping it heavily with donors, so as to introduce the Burstein shift between the emission and the absorption spectra.[9] Because of the heavy doping, the absorption edge is blue-shifted relative to the emission edge by the carrier Fermi energy. Unfortunately, the Burstein shift by itself does not seem to provide adequate transparency at room temperature, so the signal is attenuated depending on the depth of the interaction site within the semiconductor. The problem is discussed in Section 3 and its resolution in Section 4.

The transparency issue is of critical importance and we are concerned with new ways to further enhance the photon delivery to the semiconductor surface. One of the key new ideas is discussed in this paper. It is based on the extremely high radiative efficiency of high-quality direct-gap semiconductors. This means that an act of interband absorption does not finish off a scintillation photon; it merely creates a new minority carrier and then a new photon in a random direction. The efficiency of photon collection in direct-gap semiconductors is therefore limited only by parasitic processes, such as nonradiative recombination of the minority carriers and free-carrier absorption of light. The "ideal" semiconductor scintillator based on photon recycling is discussed in Section 5.

The proportionality of scintillation yield is not the only expected advantage of semiconductor scintillators. One of the major benefits of using semiconductor materials is the mature technology that enables the implementation of epitaxial photodiodes integrated on the surface of a semiconductor slab. An external receiver, like a photomultiplier, is not a viable option because of the complete internal reflection of most of the scintillating radiation. Owing to the high refractive index of semiconductors, *e.g.* $n = 3.3$ for InP, most of the scintillating photons will not escape from the semiconductor but suffer complete internal reflection. Only those photons that are incident on the semiconductor-air interface within a narrow cone $\sin\theta < 1/n$ from the interface normal have a chance to escape from the semiconductor. The escape cone accommodates only a small fraction of isotropic scintillation, $\sin^2(\theta/2) < 1/4n^2 \sim 2\%$, leading to inefficient collection.

It is therefore imperative to integrate the scintillator wafer with a photodetector that has a substantially similar or even higher refractive index in an optically tight fashion. In collaboration with Sarnoff Corporation, our group has achieved substantial progress with *epitaxial photodetectors* on InP scintillator body, implemented as ultra-low leakage *pin* diodes based on quaternary InGaAsP materials.[10] These material layers of 1.24 eV bandgap are grown epitaxially on lattice-matched InP scintillator body and are sensitive to InP scintillation at the wavelength of 0.92 μm ($h\nu = 1.35$ eV). The epitaxial diode provides nearly perfect registration efficiency of photons that have reached the heterointerface. This in turn enables three-dimensional (3D) integration of scintillator "voxels" (3D pixels) that is illustrated in Fig. 2 and discussed in the next section (Section 2).

2. Three-dimensional array of scintillator voxels: A Compton telescope

A stack of individually contacted 2D pixellated semiconductor slabs forms a 3D array of scintillator voxels, as illustrated in Fig. 2. This is the key concept of

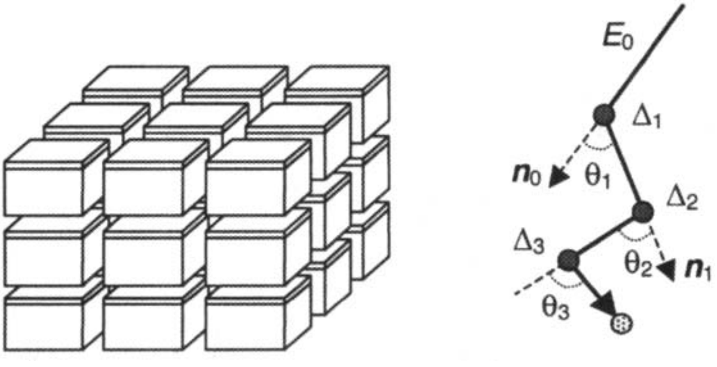


Figure 2. Schematic illustration of a 3D scintillator array. Each unit is a voxel comprising a scintillator body (white slab) and an integrated photodiode (top plate) on its surface. Advantageously, the voxel may include two photodiodes integrated on *both* surfaces in an optically tight fashion. An incident gamma photon produces a cluster of firing pixels that report their positions and the energy deposited. The information can be used to determine both the incident photon energy and the source direction. The diagram illustrates the track of a gamma photon of diminishing energy upon several successive Compton interactions producing a cluster of firing voxels.

interest to us. A gamma photon incident on such an array undergoes several Compton interactions depositing varying amounts of energy Δ_i in voxels with coordinates $\boldsymbol{r}_i = (x_i, y_i, z_i)$. The information reported enables one to determine both the incident photon energy and the direction to the source.

The exact relativistic kinematics of Compton scattering provides two equations at each interaction site, illustrated in the diagram in Fig 2:

$$\begin{aligned} \Delta_i &= E_{i-1} - E_i \\ \cos\theta_i &= 1 + 1/E_{i-1} - 1/E_i \end{aligned} \tag{1}$$

where all energies are expressed in units of electron rest energy, $m_e c^2 = 511$ keV. With three points $\boldsymbol{r}_i$ identified in the correct order ($i = 1, 2, 3$) the set of equations (1) enables one to deduce both the incident energy and the incident direction cosine from the known energies Δ_1 and Δ_2 deposited in the first two pixels and the direction θ_2 to the third pixel, as shown in Fig. 2. Direction-wise, the point source is placed on the cone of angle θ_1 about the measured direction $\boldsymbol{n}_1$. Intersection of two or three such measurements gives the direction to the source. This technique is known as the Compton telescope.[11-13]

$$E_0 = \Delta_1 + \frac{\Delta_2}{2} + \frac{1}{2}\left(\Delta_2^2 + \frac{4\Delta_2}{1 - \cos\theta_2}\right)^{1/2} \tag{2}$$

$$\cos\theta_1 = 1 + \frac{1}{E_0} - \frac{1}{E_0 - \Delta_1} \tag{3}$$

In contrast to gamma spectroscopy by conventional non-pixellated detectors,[1] a 3D array does not need to capture the entire incident gamma photon energy E_0 in order to determine E_0. However, the technique works only to the extent that one can order the events. Measurement of a fourth point in the cluster adds two equations and only one unknown and therefore can be used to determine the correct sequence. If all the measured quantities had a mathematical precision, the system of equations would be consistent only for the correct order. In practice, because both the positions and energies are measured only to within a certain margin of accuracy, the determination of the best order (referred to as the tracking algorithm or the event reconstruction) becomes a statistical procedure that minimizes inconsistency. Current tracking algorithms claim up to 70% success in correctly ordering events produced by a high-energy photon, by including in the analysis, besides the Compton kinematics, the anisotropic scattering cross-sections, expressed by the well-known[1] Klein-Nishina formula.

Semiconductor scintillators offer a tantalizing possibility of implementing a compact low-voltage Compton telescope.

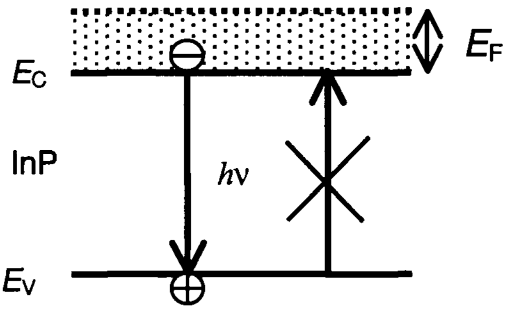


Figure 3. Schematic band diagram of a direct-gap semiconductor scintillator based on the Burstein shift (after Ref. 9). The semiconductor (InP) is doped *n*-type with degenerate doping concentration $N_D < 10^{18}$ cm^{-3}. The minority carriers (holes) are expected to recombine radiatively with electrons at the bottom of the Fermi sea. Absorption of thus emitted photons is largely suppressed.

3. Problems with scintillator based on Burstein shift

The original idea of Burstein-shift based scintillator[9] is illustrated in Fig. 3. Thermalized minority carriers (holes) are distributed near the top (within kT) of the valence band occupying a narrow range of wavenumbers $\hbar k \sim (2m_h kT)^{1/2}$. Conservation of momentum in radiative transitions requires that participating electrons have the same wavenumber and leads to an expected spectrum of emitted photons $h\nu \sim E_G + \Delta E$ of width $\Delta E \sim kT(m_h/m_e + 1)$. However, experimentally the emission spectrum $S(E)$ of heavily-doped InP is wider (of width about $E_F > \Delta E$) and it does not substantially narrow at cryogenic temperatures. This suggests that, in our heavily-doped samples, momentum is not conserved and the entire Fermi sea of electrons contributes to the spectrum, as illustrated in Fig. 4.

The Burstein shift does provide an enhanced transparency but it has not proven sufficient for the intended application. In *n*-type InP with donor concentration of $N_D = 6\text{x}10^{18}$ cm^{-3} at room temperature, the mean free path of photons averaged over the fundamental emission spectrum is about 100 μm. That is obviously not enough. However, InP has a high radiative efficiency η (over 90%) and those photons that are absorbed generate another electron-hole pair that recombines producing a new photon. We shall discuss this photon recycling process in more detail in Section 3. At this point, however, it can be noted that once photon recycling becomes a key mechanism, heavily-doped samples are no longer necessarily the best. The optimum doping concentration is set by (i) free-carrier optical absorption that obviously favors low doping; (ii) radiative efficiency, which depends on the dominant mechanism of nonradiative recombination; and (iii) the transparent fraction of the spectrum $S(E)$ that may benefit from the Burstein shift.

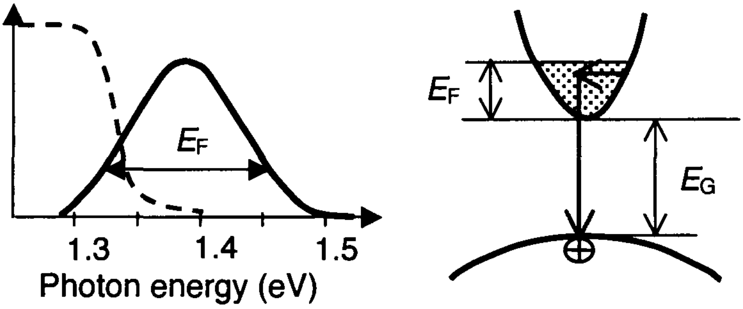


Figure 4. Schematic illustration of spectra $S(E)$ observed in heavily-doped InP (solid line). The dashed line indicates the transparency of the sample (on a logarithmic scale). The right-hand panel illustrates the radiative transitions that do not conserve quasi-momentum in heavily-doped material.

Photon recycling can deliver a reasonable fraction of the scintillating photons to the wafer surface. However, this fraction depends on the exact position of the interaction site relative the surface and this constitutes a major problem for the accurate determination of the deposited energy. One needs to distinguish the signal arising from a large energy deposited far from the photoreceiver surface from that arising from smaller energy deposited nearby.

The problem arises from the attenuation of the optical signal. If we knew the distance z of the gamma interaction event from the photoreceiver surface, one could correct for the attenuation. Recently, we came up with a method for accomplishing this, based on two-sided photodetection using epitaxial photodiodes on both sides of the InP scintillator wafer.[14] The simultaneous detection by *both* detectors of the scintillation arising from the same interaction event, allows us to determine the position of the interaction and therefore correct for attenuation.

4. Double-sided detection

Consider a scintillator slab shown in Fig. 5, with two arrays of photodetectors and two read-out circuits on the opposite sides of the slab. An essential ingredient of the proposed approach is the ability to extract the event position, characterized by a distance z the top surface, by analyzing two photoreceiver signals from the same ionization event. Our calculations indicate that it is sufficient to know the ratio of the two signals to accurately estimate of the event position. Another essential ingredient is the ability to correct for the attenuation after the position has been ascertained. Both of these capabilities can be obtained by calculations coupled with experimental validation.

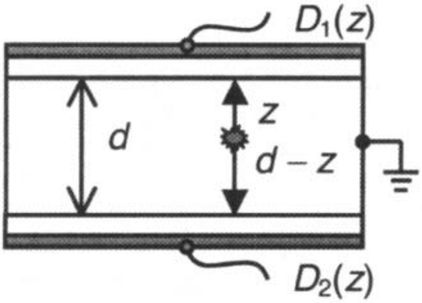


Figure 5. Schematic cross-section of an InP scintillator with two epitaxial InGaAsP photodiodes grown on both sides. Exemplarily, d = 350 μm, while the thickness of epitaxial diodes is ~2 μm (exaggerated in the figure). Lateral dimensions of the scintillator are about 1 mm.

In heavily doped InP the attenuation is due to *losses* – of photons due to free-carrier absorption and minority carriers due to nonradiative recombination. However, the ratio of attenuated signals is not sensitive to these losses, see Sect. 5.

Suppose an interaction event occurs at a distance z from the top surface (see Fig. 5). As a result, both the top and the bottom photodetectors will produce a signal. The calculated ratio $\rho(z) = D_2/D_1$ of these two signals is plotted in Fig. 6 for different scintillator doping levels. It is assumed that the photodiodes themselves are ideal and identical on both sides. Figure 6 shows that simultaneous measurement of the responses D_1 and D_2 gives sufficient information to locate the position z of the interaction event.

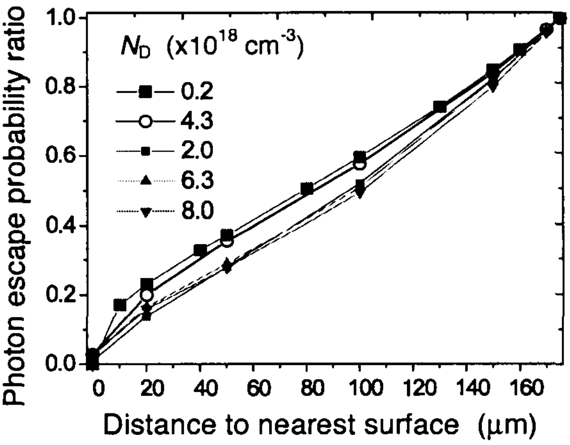


Figure 6. Calculated (see Eq. (10) in text) ratio of the two signals recorded by two surface detectors for the same radiation event for different doping levels in the scintillator body. The scintillator thickness is d = 350 μm as in Fig. 5, and the ratio is shown for $z \leq d/2$, since $\rho(z) = \rho^{-1}(d-z)$ by symmetry.

5. Scintillator based on photon recycling

In high-quality direct-gap semiconductors, most of the scintillation photons reaching the detectors' surface are not photons directly generated by the electrons and holes at the site of the gamma particle interaction, but photons that have been re-absorbed and re-emitted a multiple number of times. This phenomenon, known as photon recycling, was introduced by Dumke[15] and Moss[16] and used by Asbeck[17] to explain the long radiative lifetime in GaAs/AlGaAs heterostructures. It is often discussed in connection with high-efficiency light emitting diodes.[18] It is essential for understanding the properties of our direct-gap semiconductor scintillator.

Traditionally, an ideal scintillator requires material that is highly transparent to its own luminescence. For example, one can imagine doping the semiconductor with high efficiency radiative centers that emit below-bandgap light. It is important that the excited electron-hole pairs are efficiently transferred to the radiative center. In the case of InP, this energy transfer probability was shown[19] to be high for certain trivalent luminescent ions incorporated in the host lattice. The system InP:Yb^{3+} seems to work well at cryogenic temperatures, producing emission near 1 μm – well below the bandgap of InP. However, at room temperature, its performance is degraded by fast nonradiative de-excitation of Yb ions.[19] Other ideas for implementing *transparent* semiconductor scintillators include replacing luminescent ions by semiconductor "impregnations" of lower bandgap.[20]

In the presence of multiple photon recycling one can have another kind of ideal (*nontransparent* but lossless) scintillator, as we shall now discuss.

Let the luminescent signal comprise the energy spectrum $S(E) = GS_0(E)$ (where S_0 is normalized to unity, $\int S_0(E)dE = 1$). The signal is assumed to be generated a distance z from the detector top surface, as indicated in Fig. 5. The emitted energy is isotropic, so that the energy emitted in unit energy interval per unit solid angle is $(G/4\pi)S_0(E)$. The total energy $D_i(z)$ reaching the ith detector surface (i = 1, 2) is attenuated in a way that depends on z.

In the presence of absorption, characterized by the interband absorption coefficient $\alpha_i(E)$, the detection probability for a photon at energy E (averaged over all angles) is given by

$$p(E,z) = \int_0^\infty \exp\left[-\alpha_i(E)r\right]\frac{\cos\theta}{2r^2}\rho\, d\rho\,, \tag{4}$$

where $\rho = z\tan\theta$ and $r = z/\cos\theta$. Averaged over the emitted photon spectrum $S(E)$, the probability p_1 that the photon reaches detector 1 at $z = 0$ is given by

$$p_1(z) = \int p(E,z)S(E)\, dE\,. \tag{5}$$

We shall refer to the probability (5) and the similar probability $p_2(z) = p_1(d-z)$ that the photon generated at point z reaches detector 2 at $z = d$, as *single-pass* probabilities, because they do not include the subsequent fate (recycling) of the

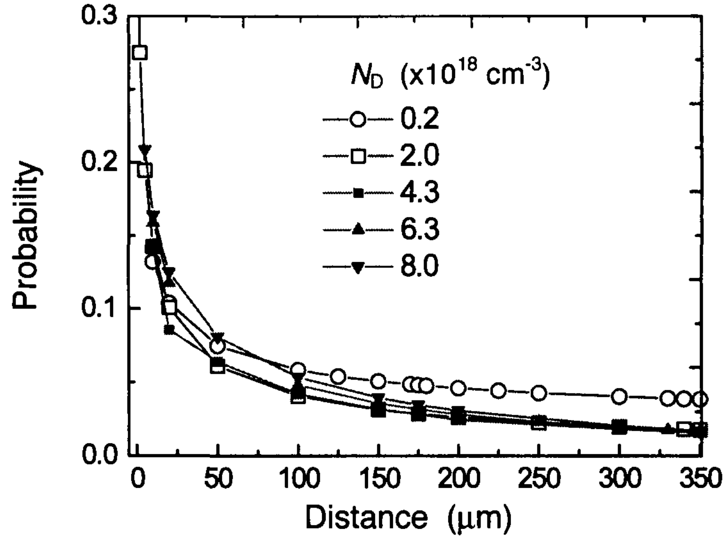


Figure 7. Single-pass photon escape probability $p_1(z)$ calculated for a 350 μm InP layer, doped *n*-type at different doping levels. The probabilities are plotted *vs.* distance from the interaction site to the integrated photodiode surface. The calculation is based on measurements of the absorption spectra in these crystals.

absorbed photon. The probability (5), calculated using experimental spectra $S(E)$ for different doping levels in InP, is shown in Fig. 7.

The single-pass probability p_{FCA} of free-carrier absorption (FCA) describes the importance of these parasitic lossy processes relative to interband absorption. It can be estimated by neglecting both p_1 and p_2, while keeping the dominant process of interband absorption,

$$p_{FCA} \approx \int \frac{\alpha_e}{\alpha_e + \alpha_i} S(E)\, dE \qquad (6)$$

where $\alpha_e(E)$ is the FCA absorption coefficient, primarily owing to intervalley electronic transitions.[19] Both α_i and α_e depend on the doping level, the former through the Burstein shift and the latter because it is directly proportional to the number of majority carriers.

Consider the basic processes in an InP scintillator with two photodiode surfaces, as in Fig. 5. The minority carrier ("hole" *h*) initially created by the ionizing interaction has the probability η (radiative efficiency) to generate a photon (of energy spectrum *S*). The generated photon can either reach detector 1 or 2 (designated by probabilities p_1 and p_2, respectively) or disappear through FCA (with probability p_{FCA}). All these probabilities depend on the position *z* (the spectral dependence can be eliminated by averaging over *S*). The combined probability $P(z) = p_1 + p_2 + p_{FCA}$ describes the likelihood of the photon loss at this stage, and the alternative, $1 - P(z)$, is the probability that a new hole is created

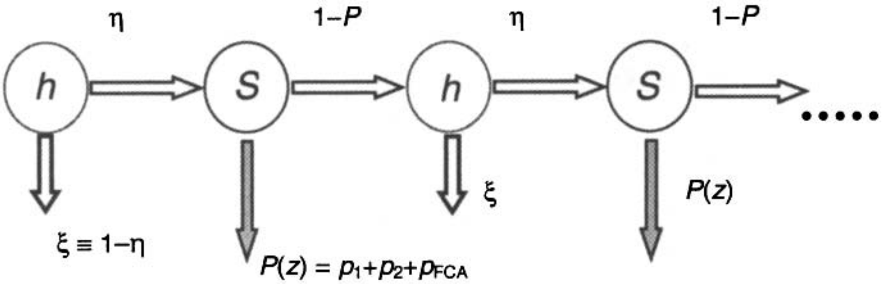


Figure 8. Schematic diagram of the basic processes in an InP scintillator with two photodiode surfaces.

through interband absorption. The cycle of hole-photon-hole transformation repeats *ad infinitum*, as illustrated in the diagram in Fig. 8.

So long as the photon recycling process continues, the minority carriers (holes) and photons are interchangeable entities. While the entity is a photon, the process can be finished off by FCA (with the probability p_{FCA}), whereas while the entity is a hole it can be ended by a nonradiative transition that occurs with the small probability $\xi = 1 - \eta$ (in our samples η ranges from 90% to 99%). Needless to say, the process can also have a happy end (photon detected). The radiative efficiency is defined in terms of the rates of radiative ($\mathrm{v_R}$) and nonradiative ($\mathrm{v_{NR}}$) transitions:

$$\eta = \frac{\mathrm{v_R}}{\mathrm{v_R} + \mathrm{v_{NR}}} \ . \tag{7}$$

The total signal in detector 1 is a sum of the single-pass contributions from different cycles of photon regeneration. As is evident from Fig. 8, it can be summed as a geometric progression, giving:

$$D_1(z) = G\eta\, p_1(z) \sum_{n=0} [\eta(1-P)]^n = \frac{G\eta\, p_1(z)}{\xi + \eta P(z)} \ . \tag{8}$$

Similarly, the total signal in detector 2 is given by

$$D_2(z) = G\,\eta\, p_2(z) \sum_{n=0} [\eta(1-P)]^n = \frac{G\,\eta\, p_2(z)}{\xi + \eta P(z)} \ . \tag{9}$$

Equations (8) and (9) express the attenuated signal in terms of single pass probabilities and photon recycling parameters η and p_{FCA}. However, the *ratio* of the two signals,

$$\rho(z) \equiv \frac{D_2(z)}{D_1(z)} = \frac{p_2(z)}{p_1(z)} \ , \tag{10}$$

depends only on the single-pass probabilities given by the spectral dependence of the *interband* absorption only, as is evident from Eqs. (4) and (5). The calculated signal ratio (10) with p_1 and p_2 as in Fig. 7, was plotted in Fig. 6 above.

It is important to note that for high photon recycling ($\eta \to 1$ and $p_{\rm FCA} \to 0$), one has an ideal scintillator in the sense that all generated photons are collected – even though the single-pass probabilities p_1 and p_2 may not be high due to interband absorption. However, the interband absorption does not "kill off" the photon because another minority carrier is created. In a sample with high η another photon will reemerge to make a new try at reaching one of the detectors. If the FCA is also low, *all photons will be eventually collected* in this way.

Let us emphasize that in such an ideal scintillator the ratio of signals will still depend on the position z of the interaction, as shown by Eq. (10). This extends the validity of the two-sided detection concept to the case of a perfect scintillator based on photon recycling. In the case of such a perfect scintillator, the value of two-sided detection is not only that it doubles the overall number of photons collected, but it also provides the position identification *not limited* to the vertical dimension of the pixel. One can have a thick pixel and still resolve the vertical position from the pre-calibrated ratio (10).

An important parameter for scintillator applications in gamma-spectroscopy is the photon collection efficiency (PCE),

$$\mathrm{PCE} \equiv [D_1(z) + D_2(z)]/G \,. \tag{11}$$

The PCE is generally less than unity, due to the above-discussed loss processes. From Eqs. (8) and (9) we find that the total PCE is given by:

$$\mathrm{PCE} = \frac{p_1(z)+p_2(z)}{\left[\left(\mathrm{v_{NR}}/\mathrm{v_R}\right)+p_{\rm FCA}(z)\right]+\left[p_1(z)+p_2(z)\right]} \,. \tag{12}$$

6. How close are we to an ideal semiconductor scintillator?

As is clear from Eq. (12), implementation of an ideal scintillator based on photon recycling requires: (a) high radiative efficiency ($\eta \to 1$ or $\mathrm{v_{NR}}/\mathrm{v_R} \ll 1$); (b) low free-carrier absorption ($p_{\rm FCA} \to 0$); and (c) reasonable transparency, expressed by nonvanishing single-pass probabilities, $p_1(z) + p_2(z) \geq 2p_1(d/2)$.

Optimization of the condition (a) depends on the nature of nonradiative recombination. The total minority-carrier recombination rate is given by

$$\mathrm{v_{TOT}} = A + Bn + Cn^2 \,, \tag{13}$$

where n is the majority-carrier concentration provided by the doping ($n = N_{\rm D}$ at room temperature). The first term A describes a nonradiative mechanism due to unintended acceptor-like impurities, whose concentration $N_{\rm A}$ (and hence $\mathrm{v_{NR}}$) is independent of $N_{\rm D}$, term B describes the radiative rate $\mathrm{v_R}$, and term C describes the Auger nonradiative recombination mechanism.

If the dominant nonradiative mechanism is A, then one would expect the optimum to be at higher $N_{\rm D}$, since the radiative rate $\mathrm{v_R} \sim N_{\rm D}$. If, on the other hand, the dominant nonradiative mechanism is due to Auger recombination ($v_{NR} \sim N_D^2$), then one should seek the optimum at lower N_D.

Our experimental results clearly favor the Auger mechanism, see Fig. 9. The solid line is a fit to the recombination rates given by Eq. (13), with the radiative (B) and Auger (C) coefficients taken from our luminescence experiments (these are close to the room-temperature date reported in the literature,[22] within their margin of accuracy) and a fitted coefficient A. The lower-doped samples have evidently higher radiative efficiency, with a maximum of about 99% at $N_D \approx 2\text{x}10^{17}$ cm^{-3}. At still lower concentrations, one can expect η to drop due to term A.

Condition (b) is clearly favored by *lower* doping concentrations, cf. Fig. 10. For $N_D \approx 2\text{x}10^{17}$ cm^{-3}, the FCA absorption is very small, $\alpha_e < 0.1$ cm^{-1}, and the probability p_{FCA} can be neglected.

Finally, condition (c) – which had been the key consideration at the start of our work – depends primarily on the positioning of the interband absorption spectrum $\alpha_i(E)$ relative to the basic luminescence emission spectrum $S(E)$, as evident from Fig. 4 and Eqs. (4) and (5). This is precisely the condition that we had sought to optimize with the Burstein shift. However, our theoretical considerations had led us to consider much lower concentration ranges, where the Burstein shift is inoperative. The experimentally observed luminescence spectra show that the low-doped (non-degenerate) sample with $N_D = 2\text{x}10^{17}$ cm^{-3} is about ten times brighter than our best degenerately doped samples. In the low-doped sample both the quantum efficiency becomes high (due to the absence of Auger recombination) and the residual (free-carrier) absorption becomes very low. The feared drawback of lower-doped samples, the tight overlap of the absorption and emission bands, is more than compensated by high photon recycling.

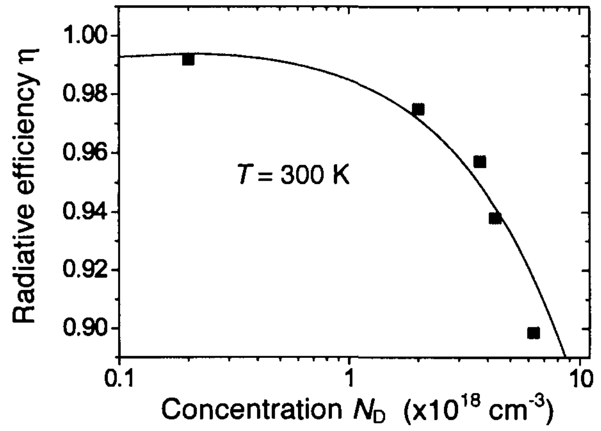


Figure 9. Room-temperature radiative efficiency as function of majority-carrier concentration provided by the doping N_D. The solid line describes a fit to Eq. (13) with $A = 2\times10^5$ s^{-1}, $B = 1.9\times10^{-10}$ cm^3/s, and $C = 1.5\times10^{-30}$ cm^6/s.

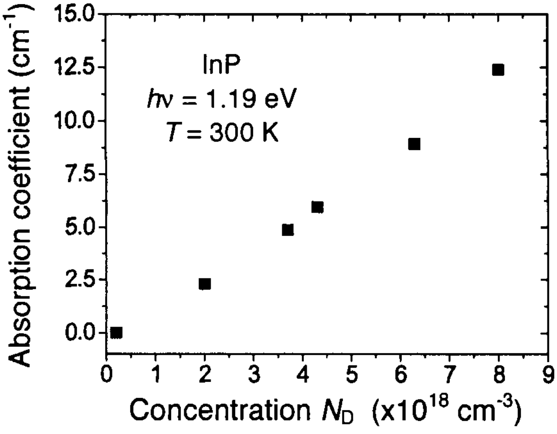


Figure 10. Free-carrier absorption coefficient α_e *vs.* N_D measured in our InP samples. At the lowest concentration $N_D \approx 2\times10^{17}$ cm^{-3}, one has $\alpha_e < 0.1$ cm^{-1}.

Once the conditions (a–c) are satisfied, the photon collection efficiency predicted by Eq. (12) is very high indeed, even when single-pass probabilities are low. With negligible free-carrier absorption ($p_{FCA} \to 0$) and the high radiative efficiency ($v_{NR}/v_R = 0.01$) featured by our low-doped InP samples, the PCE = 0.95 assuming $p_1(z) + p_2(z) \approx 0.2$, as estimated from our earlier experimental data. Even for still lower single-pass probabilities, such as $2p_1(d/2) \approx 0.1$, estimated theoretically (see Fig. 7), one has PCE = 0.91. It should be noted that the theoretical estimate presented here gives a lower estimate for the PCE. A more detailed theoretical analysis, now in progress, will allow for (i) finite size of the initial excitation region, and (ii) expansion of the excitation region in the process of recycling. Both effects are expected to increase the PCE estimate. Inclusion of (i) is reasonably straightforward, provided one knows the spatial distribution $H(\boldsymbol{r})$ of holes generated by the gamma photon. One would then simply average Eqs. (8) and (9) over $H(\boldsymbol{r})$. Evaluation of $H(\boldsymbol{r})$ is not a simple matter, however.[23,24]

Inclusion of the effect (ii) is more subtle, especially for the lower-doped samples. Our derivation of Eqs. (8) and (9) assumes that every act of recycling occurs at the same place z where the initial interaction occurred, and therefore the same probabilities $p_1(z)$ and $p_2(z)$ appear at all stages of the recycling, see Fig. 8. This has reduced the summation of an infinite series to a geometric progression and allowed us to obtain the result in a closed form. In reality, however, there is a transport of holes in photon recycling, which has the nature of a random walk. Recombination of a hole at a position $\boldsymbol{r}$ is accompanied by the emergence of another hole at a $\boldsymbol{r} + \boldsymbol{\ell}$, where $\boldsymbol{\ell}$ is a random vector, characterizing the free flight of an emitted photon. In a heavily doped sample this random walk reduces to diffusion with a diffusion coefficient $\sim|\boldsymbol{\ell}|^2/\tau$, where τ is the radiative recombination time. For a sample with $N_D = 6\times10^{18}$ cm^{-3} we estimate $|\ell|$ to be of order 100 μm. For lightly-doped samples, however, the diffusion approximation breaks down because of the extreme transparency in the long wavelength wing of the spectrum $S(E)$, where $|\ell|$ is limited by FCA only. For the sample with $N_D \approx 2\times10^{17}$ cm^{-3}, where $\alpha_e < 0.1$ cm^{-1}, one has $\alpha_e d \ll 1$, and nothing should change if we set $\alpha_e = 0$, which corresponds to diverging free flight $|\ell| \to \infty$. Random walks with a distribution of unit flights so long-tailed that their second moment diverges are known as Lévy flights.[25] An interesting example of optical Lévy flights was recently discussed by Barthelemy *et al.*[26] Such random walks with divergent average step length lead to the so-called *super-diffusion*, where the average squared displacement increases superlinearly with time. As subtle as this may be, the super-diffusive behavior is likely to be of importance for estimating the photon collection efficiency. By a crude analogy with averaging over $H(\boldsymbol{r})$, we expect that the single-pass probabilities in Eq. (12) will be replaced by larger "effective" values $p'_i(z) > p_i(z)$, further enhancing the PCE estimate. The enhancement is hard to predict at this time and it requires further studies. The ultimate aim of these studies is an accurate estimate of the optimum scintillator layer thickness for best energy resolution.

7. Conclusions

As discussed in the introduction, the semiconductor scintillator does not suffer from the non-proportionality effects and its energy resolution is fundamentally limited by energy branching in the same way as it is for semiconductor diodes. However, high photon collection efficiency is an obvious prerequisite for the good energy resolution. We have shown that a semiconductor scintillator based on photon recycling is capable of delivering very high PCE, even though the material is not transparent in the usual sense to its own luminescence. In order to take advantage of photon recycling, double-sided integrated photodiodes appear necessary. We can ultimately expect the energy resolution of semiconductor scintillators to approach that of a diode detector implemented in the same material.

Acknowledgments

This book chapter would not be possible without the contributions of the radiation detector research group at Stony Brook University, including Alex Kastalsky, Michael Gouzman, Nadia Lifshitz, Oleg Semyonov, Milutin Stanacevic, and Vladislav Kuzminsky, and graduate students William Cheng, Vladimir Smagin and Zhichao Chen. We would also like to acknowledge important contributions of our friends and collaborators at Sarnoff Corporation, Joseph Abeles, Winston Chan and Zane Shellenbarger.

This work was supported by the Domestic Nuclear Detection Office (DNDO) of the Department of Homeland Security, by the Defense Threat Reduction Agency

(DTRA) through its basic research program, and by the New York State Office of Science, Technology and Academic Research (NYSTAR) through the Center for Advanced Sensor Technology (Sensor CAT) at Stony Brook.

References

1. G. F. Knoll, *Radiation Detection and Measurement*, 3rd ed., New York: Wiley, 2000.
2. R. Hofstadter, "Alkali halide scintillation counters," *Phys. Rev.* **74**, 100 (1948).
3. C. L. Melcher, "Perspectives on the future development of new scintillators," *Nucl. Instrum. Methods Phys. Res. A* **537**, 6 (2005).
4. P. Dorenbos, "Light output and energy resolution of Ce^{3+}-doped scintillators," *Nucl. Instrum. Methods Phys. Res. A* **486**, 208 (2002).
5. A. N. Vasil'ev, "From luminescence nonlinearity to scintillation non-proportionality," *IEEE Trans. Nucl. Sci.* **55**, 1054 (2008).
6. G. Bizarri, W. W. Moses, J. Singh, A. N. Vasil'ev, and R. T. Williams, "An analytical model of nonproportional scintillator light yield in terms of recombination rates," *J. Appl. Phys.* **105**, 044507 (2009).
7. G. Bizarri, W. W. Moses, J. Singh, A.N. Vasil'ev, and R. T. Williams, "The role of different linear and nonlinear channels of relaxation in scintillator non-proportionality," *J. Luminescence,* in press (2009).
8. S. A. Payne, N. J. Cherepy, G. Hull, J. D. Valentine, W. W. Moses, and W.-S. Choong, "Nonproportionality of scintillator detectors: Theory and experiment," *IEEE Trans. Nucl. Sci.* **56**, 2506 (2009).
9. A. Kastalsky, S. Luryi, and B. Spivak, "Semiconductor high-energy radiation scintillation detector," *Nucl. Instrum. Methods Phys. Res. A* **565**, 650 (2006).
10. S. Luryi, A. Kastalsky, M. Gouzman, N. Lifshitz, O. Semyonov, M. Stanacevic, A. Subashiev, V. Kuzminsky, W. Cheng, V. Smagin, Z. Chen, J. H. Abeles, W. K. Chan, and Z. A. Shellenbarger, "Epitaxial InGaAsP/InP photodiode for registration of InP scintillation," to be published.
11. E. Aprile, A. Bolotnikov, D. Chen, and R. Mukherjee, "A Monte Carlo analysis of the liquid xenon TPC as gamma-ray telescope," *Nucl. Instrum. Methods Phys. Res. A* **327**, pp. 216-221 (1993).
12. J. D. Kurfess, W. N. Johnson, R. A. Kroeger, and B. F. Phlips, "Considerations for the next Compton telescope mission," *Proc. 5th Compton Symp.*, Portsmouth, NH (1999).
13. S. E. Boggs and P. Jean, "Event reconstruction in high resolution Compton telescopes," *Astron. Astrophys. Suppl. Ser.* **145**, 311 (2000).
14. J. H. Abeles and S. Luryi, "Slab scintillator with integrated two-sided photoreceiver," U.S. prov. patent application, Ser. No. 61/216,572 (2009).
15. W. P. Dumke, "Spontaneous radiative recombination in semiconductors," *Phys. Rev.* **105**, 139 (1957).
16. T. S. Moss, "Theory of the spectral distribution of recombination radiation from InSb," *Proc. Phys. Soc. B* **70**, 247 (1957).

17. P. Asbeck, "Self-absorption effects on the radiative lifetime in GaAs/AlGaAs heterostructures," *J. Appl. Phys.* **48**, 820 (1977).
18. See, for example, R. Windisch, C. Rooman, B. Dutta, A. Knobloch, G. Borghs, G. H. Döhler, and P. Heremans, "Light-extraction mechanisms in high-efficiency surface-textured light-emitting diodes," *IEEE J. Selected Topics Quantum Electronics* **8**, 248 (2002).
19. A. Taguchi and K. Takahei, "Trap level characteristics of rare-earth luminescence centers in III–V semiconductors," *J. Appl. Phys.* **79**, 4330 (1996).
20. S. Luryi, "Impregnated semiconductor scintillator," *Intern. J. High Speed Electronics Syst.* **18**, 973 (2008).
21. W. P. Dumke, M. R. Lorenz, and G. D. Pettit, "Intra- and interband free-carrier absorption and the fundamental absorption edge in *n*-type InP," *Phys. Rev. B* **1**, 4668 (1970).
22. See the electronic archive on semiconductor materials, characteristics and properties at *http://www.ioffe.rssi.ru/SVA/NSM/*, maintained by the Ioffe Physico-Technical Institute.
23. E. Gros d'Aillon, J. Tabary, A. Glière, and L. Verger, "Charge sharing on monolithic CdZnTe gamma-ray detectors: A simulation study," *Nucl. Instrum. Methods Phys. Res. A* **563**, 124 (2006).
24. J. Hayward and D. Wehe, "Incomplete charge collection in an HPGe double-sided strip detector," *Nucl. Instrum. Methods Phys. Res. A* **586**, 215 (2008).
25. See, for example, R. Metzler, A. V. Chechkin, V. Yu. Gonchar, and J. Klafter, "Some fundamental aspects of Lévy flights," *Chaos, Solitons, Fractals* **34**, 129 (2007).
26. P. Barthelemy, J. Bertolotti, and D. S. Wiersma, "A Lévy flight for light," *Nature* **453**, 495 (2008).

Semiconductor Gamma Radiation Detectors: Band Structure Effects in Energy Resolution

Arsen Subashiev and Serge Luryi
Dept. of Electrical and Computer Engineering, SUNY–Stony Brook
Stony Brook, NY 11794-2350, U.S.A.

1. Introduction

Energy spectroscopy in semiconductor γ-detectors is based on registration of the total number of electron-hole (e-h) pairs produced by a single γ-photon after a cascade of various processes. These include Compton scattering events, photo-absorption, deep core level excitation, core vacancy relaxation, emission of plasmons by the secondary electrons (holes), plasmon decay into e-h pairs and impact ionization events. At each stage, the cascade is accompanied by sequential energy branching between secondary particles, resulting in a nearly random energy distribution in a cloud of secondary electrons and holes in the final state. The branching is terminated when the energy of a secondary electron or hole is below the impact ionization threshold. Importantly, to a very high precision and in a broad energy range, the number of created pairs N is just proportional to the initial energy $E = h\nu$ of the γ-photon, $N = E/\varepsilon$, where ε is the average excitation energy of a single pair (referred to as the pair excitation energy). For semiconductor materials, the pair excitation energy ε is about $3E_G$, where E_G is the band gap energy and is typically about $2E_M$, *i.e.* twice the minimal energy required for an electron to produce a pair by impact ionization in a process with momentum conservation. This means that electrons and holes possess a large average residual kinetic energy after cascade termination.

Due to the almost random nature of the cascade energy branching, the pair number N fluctuates from one event to the other. Therefore, the theoretical limit for the γ-detector energy resolution depends on the variance of the registered energy E, *i.e.* $\langle\delta E^2\rangle = \langle\delta N^2\rangle\varepsilon^2$. The variance differs from the Poissonian statistics by the so-called Fano factor F,[1] so that $\langle\delta N^2\rangle = F\langle N\rangle$. Experimental data for F are most reliably determined for Ge and Si, where they are close to 0.1. For many years the statistical limitations of energy resolution expressed by the factor F were less important than those due to electronic noise and the effects of incomplete charge collection. However, owing to recent progress in electronics[2] and the development of pixel detectors,[3] the Fano factor contribution to energy resolution has become tangible. Moreover, there is current research interest in a number of novel detectors, based on new energy-measurement principles (such as the Compton telescope[4,5]), as well as new semiconductor detector materials.[6]

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

Therefore, a critical evaluation of the Fano factor theory is relevant the search for materials with improved energy resolution.

Theoretical models used for Fano factor calculations have been based on widely varying assumptions, often ignoring the important details of the energy branching process.[7,8] Therefore, the confidence in these models is questionable even though they routinely produce estimates of F in the range 0.06 to 0.14, close to the experimental results for a number of semiconductors.[7] In this chapter, we review some of the popular models used and discuss the most subtle aspect of energy branching, namely the pair correlations in the final-state energy distribution. We show that these correlations need being taken into account and, moreover, that they can be used in the search for materials with a reduced Fano factor and hence improved intrinsic energy resolution.

2. Final state models

The pair production in semiconductor γ-detectors occurs by an ionization cascade that goes through a number of intermediate stages, where the energies of secondary electrons or holes are much larger than the ionization energy. The relative energy loss at these stages is typically much smaller than the energy spread in the final state, as has been shown by Monte-Carlo modeling,[9,10] using fairly well-known cross-sections of elementary processes. For secondary high-energy electrons, the dominant energy-loss process for high-energy electrons is the emission of plasmons (collective valence-band excitations).[7,11] The density of valence-band electrons is almost the same for all tetrahedral semiconductors and the plasmon energy is close to 16 eV for Ge, Si, and III-V semiconductors. De-excitation of the deep core state vacancies goes through fluorescence or Auger-like Coster-Kronig transitions, also creating intermediate electrons and holes, spread over a wide energy interval in the conduction and valence bands, respectively.

The final stage of the energy branching is governed by a competition between impact ionization and phonon emission. The energy conservation in impact ionization for an electron with initial energy E_e reads $E_e = E_G + E_{1e} + E_{2e} + E_h$, where E_{1e}, E_{2e}, and E_h are the kinetic energies of carriers (in this example, two electrons and a hole, respectively). Similar relation holds for impact initiated by a hole. After the bandgap energy E_G is subtracted, the residual energy is distributed between 3 particles. The cascade process is repeated until the final energy of a secondary particle is smaller than the impact ionization threshold or until subsequent phonon emission becomes more probable. Termination of the cascade fixes the number N of pairs, since the excess energy dissipates without changing the particle number.

The most popular statistical approach to the Fano factor evaluation is based on the analysis of the energy distribution in the system of e-h pairs and phonons in the *final state*, *i.e.* at the stage when further creation of e-h pairs is impossible.

Consider the simplest case where phonon emission is ignored. Then, for a particular realization of the cascade process, the initial energy E is distributed between N pairs, $E = \Sigma E_i$, where E_i is the energy of the ith pair. In this distribution, the sum E is fixed, whereas the number of terms N is the statistical variable. It is instructive, however, to begin with the opposite situation, with a sum of a fixed number N of independent random variables E_i, whose distribution is characterized by a mean $\varepsilon = \langle E_i \rangle$ and a variance $\langle \delta\varepsilon^2 \rangle = \langle E_i^2 \rangle - \varepsilon^2$.

In this case, according to the central limit theorem, the sum E will be normally distributed about its mean value $N\varepsilon$:

$$P(E) = C_N \exp\left[-\frac{(E - N\varepsilon)^2}{2N\langle \delta\varepsilon^2 \rangle}\right], \tag{1}$$

where C_N is a normalization constant. For the case we are interested in, when the total (initial) energy is fixed for all events, while the number of pairs fluctuates from one realization of energy branching to another, we can reinterpret Eq. (1) to provide the probability distribution for N:

$$P(N) = C \exp\left[-\frac{(N - \langle N \rangle)^2 \varepsilon^2}{2\langle N \rangle \langle \delta\varepsilon^2 \rangle}\right], \tag{2}$$

where C is another normalization constant; we have replaced N by $\langle N \rangle$ in the denominator since their difference is small. It is evident from Eq. (2) that the variance in the particle number is not Poissonian and that the Fano factor equals

$$F = \frac{\langle \delta\varepsilon^2 \rangle}{\varepsilon^2}. \tag{3}$$

The above argument leading to the widely used expression (3) for the Fano factor can be illustrated by the so-called "shot-glass" model.[12] In this model, the initial energy E deposited in the semiconductor is taken analogous to the volume of water in a bathtub, while the e-h pair creation energy is represented by a volume of a shot glass with the help of which the tub is emptied sequentially. The shots are thought of being filled differently in a random fashion with a known mean and variance. The total number of shots used to empty the tub will fluctuate from one realization to another and one can apply the above reasoning leading to Eq. (3). The shot-glass model is an example of the class of uncorrelated models, called the *final-state models*. It assumes that the final state is formed as a composition of uncorrelated random events described by the parameters (mean and variance) of the single-event statistics, with no correlations between the events. The apparent non-Poissonian statistics in the final-state models results from one single constraint: the fixed total initial energy. As a result, Eq. (3) readily provides smaller than unity values for the Fano factor.

Equation (3) can be generalized to include "inelastic energy losses", such as phonon emission. The energy loss associated with phonon emission is much lower than the ionization threshold. Phonon emission rate is also much lower than that of impact ionization, except very near the threshold where the ionization rate is much reduced. Near the threshold, the emission of a single phonon can make the next impact ionization impossible. The apparent shift of the ionization threshold can be accounted for by an adjusted value of ε. In the final state models these events are assumed independent, so that fluctuations of the phonon energy losses are additive to those in E_i. The modified Eq. (3) is then of the form

$$F = \frac{\left\langle \delta\varepsilon^2 \right\rangle + (\hbar\omega)^2 \rho_{ph}}{\varepsilon^2}, \tag{4}$$

where ρ_{ph} is the average number of phonons emitted per e-h pair (created before the cascade termination). The number of emitted phonons is assumed to have Poissonian statistics. The average final-state pair energy is $\varepsilon = E_G + \langle K_e \rangle + \langle K_h \rangle$, where K_e and K_h are the kinetic energies of electrons and holes, distributed in a range from zero to the impact ionization threshold E_{th}. The number of phonons emitted before cascade termination is small and so are the corresponding energy losses, $\hbar\omega\rho_{ph} << N\varepsilon$. The losses coming from other inelastic processes, such as fluorescence of deep-level holes, bremsstrahlung emission with some radiation escape from the crystal, *etc.* are also similarly small.

In order to use Eq. (4), one needs to know or assume the energy distribution in the final state. The simplest assumption is that final-state electrons are homogeneously distributed in momentum space (*k*-space) with a parabolic energy band spectrum.[13] This implies a distribution proportional to the band density of states (DOS) and cut off at E_{th}. For this distribution, the average pair kinetic energy $\langle K_e \rangle + \langle K_h \rangle = 1.2E_M$, where $E_M = 1.5E_G$ is a maximum allowed energy equal to the ionization energy threshold. Then, the average energy needed for a pair production comes out to be $\varepsilon = E_G + 1.2E_M = 2.8E_G$, close to experimental data. On the other hand, calculations of $\langle \delta\varepsilon^2 \rangle$ (with the same assumption of homogeneously distributed final-state particles in *k*-space up to the energy of $1.5E_G$) give $2\langle (\delta K_i)^2 \rangle = 9/2 \times 0.07E_G^2 = 0.315E_G^2$. This yields the electronic contribution to the Fano factor to be ~0.04, far below the experimental data. To obtain a Fano factor that is closer to experiment (at least a factor of two larger), one either has to assume an unrealistically large phonon contribution or find other channels of energy loss. Thus, in order to obtain good agreement with the experimental data for Ge, the author of a widely cited paper[14] assumed the energies of electrons and holes spread over half of the conduction and valence band, respectively, while the densities of states were assumed constant. The author of a well-known book[2] assumes that the total number $n = (E - NE_G)/\hbar\omega$ of phonons emitted in the pair production fluctuates according to the Poisson statistics. This assumption (also yielding $F < 1$) has no sound justification and leads to a strong dependence of the Fano factor on the phonon energy. For most semiconductors it overestimates the value of F (compared to experiment). Other authors[15] calculated the average number of phonons emitted before the cascade termination and thus estimated the average phonon energy loss as a function of initial electronic energy, which represents the second term in Eq. (4). However, their approximate approach used to evaluate the energy distribution function (now extending far beyond the impact ionization threshold) also required an adjustment to the bandgap energy E_G to obtain reasonable values of Fano factor and pair excitation energy.

Within the final-state model framework, the phonon losses are statistically independent and should be treated as simply additive in calculations of the average excitation energy and the Fano factor. If only a few phonons are emitted per pair, the phonon contribution is rather negligible. Microscopic consideration of energy branching readily suggests that the main effect of phonon emission is a strong shift of the impact ionization threshold, and thus the increase of the width of the kinetic energy distribution. This affects both the average excitation energy and the Fano factor, even when the average number of emitted phonons is small.

This review of the final-state model approach is prompted by a recent paper,[12] which asserts that the model "provides a convenient framework for and point of departure for analyzing the essential statistical features of microscopic theories".

It is clear that usefulness of final-state models is limited by the need to make assumptions about the final state distribution that go beyond the model itself. As a result, these models may often be misleading since the key assumptions, such as the role of phonon losses, may not be justified.

Nevertheless, it is important to examine the final-state model approach *in principle*, leaving aside all uncertainties in the additional assumptions. This has been done[16] on the basis of a sequential energy branching model that can be solved exactly. In this, the so-called random parking model, one can calculate exactly both the Fano factor and its representation by the final-state model. The exact result sheds light on the principal drawback of the latter.

The problem with all final-state models is that they postulate statistically independent final-state pair energies. This leads to the Gaussian distribution (1) and further to Eq. (3). However, the kinetics of an ionization cascade produces a *correlated* energy distribution in the final state, so that for a pair of electrons in the final state with energies E_i and E_j one has $\langle (E_i - \varepsilon)(E_j - \varepsilon) \rangle \neq \delta_{ij}(\langle E_i \rangle^2 - \varepsilon^2)$. Correlations between different terms in $E = \Sigma E_i$ invalidate Eqs. (1) and (3), whose derivation was based on the central limit theorem.

The importance of correlations can be easily illustrated in the shot-glass model: imagine that the man with the shot-glass watches what he is doing and compensates for an underfilled shot by following it with a larger than average shot, so that two successive glasses together scoop the same amount of water. Evidently, the fluctuations will be strongly reduced compared to predictions of Eq. (3). One can observe that this type of correlations is typical for energy branching: *e.g.* the division of initial energy between two particles produces a state where small energy of one particle is paired with large energy of the other. We shall discuss these correlations further and show that they are of major importance at the final stage of the energy branching.

Another relevant observation that follows from our exact model analysis[16] is the importance of the *shape* of final energy distribution. It turns out that the exact one-particle distribution has a narrow peak at small energies. Therefore, combined with exact statistics, Eq. (3) will always provide a lower bound estimate for F.

3. Kinetics of energy branching: One-particle distribution function

For a realistic band structure, phonon spectrum, and scattering matrix elements, the calculations are necessarily numerical. To analyze qualitative features, we discuss a kinetic approach to the distribution function and present analytical results for a simplified model, where the valence band is so narrow, that one can neglect the kinetic energy of created hole. In this case, the energy at the every stage of the cascade is distributed between the two secondary electrons. This postulate and some of its generalizations permit of an analytical solution.[14,15,17]

Correlations in sequential energy branching are fully included in the so-called "crazy carpentry" (CC) model due to van Roosbroeck.[17] It illustrates the branching process by random cutting of a wooden board into two peaces with the removal of a unit-length board from one piece (if its length is less than unity, the piece is discarded). The removed unit boards are stored and the process is repeated discarding all lengths less than unity until the entire board is used up. The initial board length is associated with the incident energy and the number of stored unit-length boards with the number of e-h pairs produced in the sequential energy branching. The discarded pieces represent the residual kinetic energies of electrons and holes in the final state.

An alternative but perhaps more transparent illustration of sequential energy branching is the so-called "random parking problem" (RPP). It is equivalent to the CC model[17] and leads to the same equations and analytical results. In this model one studies the distribution of cars of unit size in a long linear parking lot, assuming that they arrive sequentially in time at random position in the lot. A car is rejected if it lands at an occupied space. The process is terminated when the so-called "jamming limit" is achieved, *i.e.* when all remaining gaps between the cars do not exceed unit length. The energy branching can be considered in terms of the RPP if one identifies the incident particle kinetic energy with the parking lot length and the pair creation energy with the car size. The quantities of interest are the filling factor (average number of cars per unit length), equivalent to the average pair excitation energy, and the fluctuation of the number of cars in different realizations that can be represented by the Fano factor. A number of generalized RPP problems have been studied, including the model of cars shrinking after they are parked[18] that describes the case when the minimum pair creation energy is larger then the band gap (the energy removed from the system in pair creation), and the two-size parking problem[19] that models the semiconductor with several minima in the conduction band (resulting in several impact ionization thresholds).

The key distinction of the CC and RPP models from the final-state models is the account of *sequential* energy branching that leads to correlations in the final distribution. In contrast, the final state models sample the energy repeatedly in an uncorrelated fashion.[12] Nevertheless, application of the CC and RPP models to sequential energy branching in radiation detectors was viewed with some suspicion. The early calculations (including Monte-Carlo modeling) of the excitation energy and the Fano factor had led to the belief that reliable results require an almost infinitely long parking lot (corresponding to very high initial energy, at least $E_0 > 100E_G$). At these high energies, processes other than impact ionization (*e.g.* plasmon emission, deep core electron excitation and so on) make the energy distribution between the secondaries highly inhomogeneous. This would render the results meaningless.[12] However, we have shown[18,19] that the recursion division process saturates at a much earlier stage of initial energies, $E_0 \approx 7\text{–}10E_G$, after which the results become insensitive to the initial energy distribution. It was also shown[18,19] that the recursive and the kinetic approaches are equivalent in the limit of infinitely high initial energy.

In this section, we consider kinetic evolution of the one-particle distribution function in the RPP model of sequential energy branching with a *finite* initial energy. Having obtained the exact single-particle distribution in the jamming limit, we then calculate the average excitation energy and the Fano factor using Eq. (3).

The kinetic equation for the electron energy distribution function $f(E)$ can be written in the form

$$\frac{\partial f(E,t)}{\partial t} = -R(E)f(E,t) + d\int_{E+E_t}^{E_0} dE'\rho(E')P(E',E)f(E',t) \ , \tag{5}$$

where $R(E)$ is the impact ionization rate for an electron with initial energy E and $P(E',E)$ is the probability of an impact ionization process with initial energy E' and final electron energy E. The DOS function $\rho(E)$ is for initial states. The second term of Eq. (5) describes electron ionization resulting in a new electron in the state with energy E.

The factor d represents the degeneracy of the branching by impact ionization. For a narrow valence band, there are $d = 2$ final states with the same final electron energy; for a model with symmetrical conduction and valence bands, $d = 3$. In the RPP model, the function $f(E, t)$ represents the distribution of gaps between cars at the moment t. The same new spacing can evidently be obtained either on the right or on the left side of the newly arrived car.

Kinetic equation (5) describes the evolution of the energy distribution in the branching cascade and evidently, it does not conserve the number of particles. However, it is consistent with the total energy conservation law in a form

$$\int_0^\infty \left(E + \frac{E_{\text{th}}}{d-1}\right) f(E,t)\rho(E)dE = E_0 + \frac{E_{\text{th}}}{d-1} \ . \tag{6}$$

Multiplying Eq. (5) by $[E + E_{\text{th}}/(d-1)]\rho(E)$, integrating it over energy and using Eq. (6), one can find $R(E)$ for any given $P(E', E)$. Equation (5) should be solved with the initial condition $f(E, t = 0) = C\delta(E-E_0)$ for the branching process initiated by a high-energy electron created by the initial Compton event; the constant C can be easily found from Eq. (6): $C = \rho^{-1}(E_0)$. Next, the number of pairs is given by

$$N_p = \frac{1}{d-1}\left[\int_0^\infty f(E,t)\rho(E)dE - 1\right] , \tag{7}$$

so that the pair excitation energy equals

$$\varepsilon = N_{\mathrm{p}}^{-1}\left(E_0 + \frac{E_{\mathrm{th}}}{d-1}\right) . \tag{8}$$

Equation (8) may seem to be at variance with the former definition of the pair excitation energy. But that definition referred to the $N_{\mathrm{p}} >> 1$ limit, where the additional term in Eq. (8) is negligible. For finite and especially the intermediate stages of energy branching with E_0 in the range of 3–5E_{th}, this correction ensures the linear growth with energy of the number of created pairs.

To study the kinetics of the energy branching, it is convenient to use the Laplace transformation in time,

$$F(E,p) = \int_0^{\infty} e^{-pt} f(E,t)dt . \tag{9}$$

Applied to Eq. (5) this gives

$$pF(E,p) = -R(E-E_{\mathrm{th}})F(E,P) + d \int_{E+E_{\mathrm{th}}}^{\infty} P(E',E)F(E',p)\rho(E')dE' + f(E,0) . \tag{10}$$

Two important observations are in order. First, Eq. (10) can be used for any initial energy E_0 of the particle initiating the cascade process. Second, the Laplace transform function $F(E,p)$ is expressed through $F(E+E_{\mathrm{th}},p)$ and therefore it can be found recursively, starting from the energy range $E_0 - E_{\mathrm{th}}$ and then going down to lower energies. For the case of the narrow valence band, the temporal evolution of normalized energy distributions,

$$f^*(E,t) = \frac{f(E,t)}{E_0 + E_{\mathrm{th}}/(d-1)} , \tag{11}$$

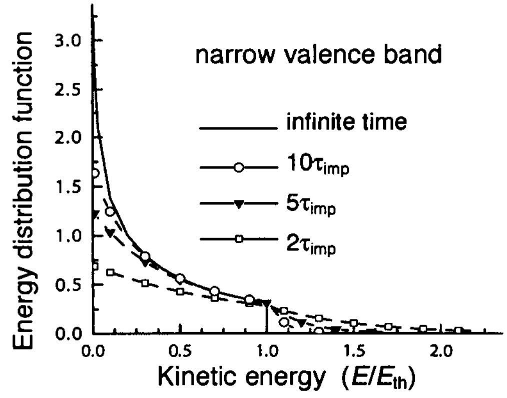


Figure 1. Evolution of the electron energy distribution function in case of a narrow valence band from initial state with one particle with energy $20E_{\mathrm{th}}$ (τ_{imp} denotes the impact ionization time).

is shown in Fig. 1. Note the rapid decay of the high-energy tail of the distribution and growth of the peak near zero. This peak in the population of low-energy states leads to the underestimation of the Fano factor by the final-state model.

The final distribution can be easily obtained using the Heaviside expansion theorem for the Laplace transforms, which gives

$$f(E,t \to \infty) = d \int_{E+E_{\text{th}}}^{\infty} P(E')F(E',0)\rho(E)dE' , \quad 0 < E < E_{\text{th}} . \tag{12}$$

Here $F(E',0)$ is the solution of Eq. (12) at $p = 0$, viz.

$$F(E,0) = R(E - E_{\text{th}})^{-1} d \int_{E+E_{th}}^{\infty} P(E')F(E',0)\rho(E')dE' + f(E,0) \tag{13}$$

Next, we trace the convergence of the distribution to its limiting form with the increase of E_0. For a narrow valence band, the distribution function for $E_0 \to \infty$ is known from the solution of the RPP with an infinitely large parking lot:

$$f^*(E,t \to \infty) = 2\int_0^{\infty} se^{-Es-2I(s)}ds , \tag{14}$$

where $f^*(E) = f(E)/E_0$ (with $E_0 \to \infty$) is the normalized distribution function, and

$$I(s) = \int_0^{s} \frac{1-e^{-u}}{u} du \quad . \tag{15}$$

The normalized distribution functions $f^*(E,t \to \infty)$ calculated for several initial energies E_0 are shown in Fig. 2.

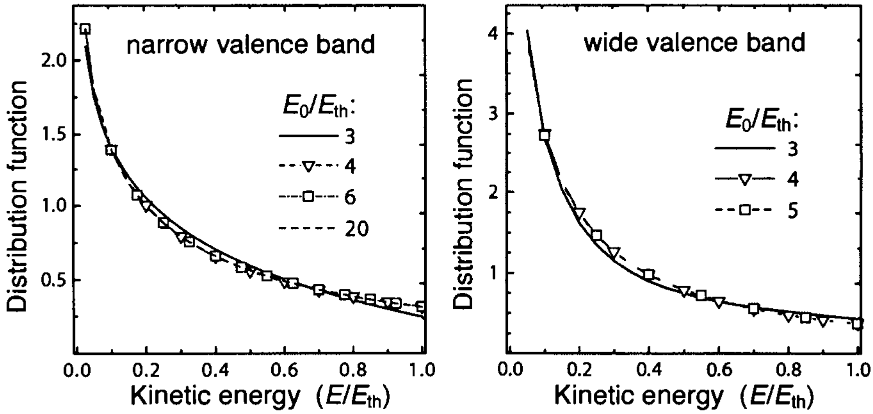


Figure 2. Average (over realizations) one-particle kinetic energy distribution functions after the energy branching termination. Different lines correspond to different initial particle energy E_0 in units of E_{th}.

As the initial energy increases, one can observe a very rapid convergence of the distribution to its final form corresponding to $E_0 \rightarrow \infty$. To within 1%, the results become indistinguishable already at $E_0 = 4E_{th}$. The obtained distribution functions are precise enough for the calculations of the average kinetic energy and the number of pairs in the final state (and thus both the pair excitation energy and its variance).

Similar analysis can be done for a finite-width valence band (see the right panel of Fig. 2). The rapid convergence is again in evidence. This convergence, found to be a general property of the branching process,[18,19] leads to an important conclusion: the details of energy branching at the initial and intermediate stages are not important for the final state energy distribution, since they do not produce an excessive number of particles with small energy. Further averaging over the distribution of secondaries in the intermediate stages has no influence on the final distribution and hence on the energy resolution.

This conclusion is very important for the energy branching problem and deserves to be further illustrated within the RPP model and formulated more accurately. Suppose at the initial stages of filling the parking line, the cars are trying to observe some rules, say, parking regularly at a distance $N_1 E_{th}$ from each other. Our conclusion is that the effect of this strong ordering on the final distribution function will be precisely described as

$$f_C(E) = f(E)\left(1 + \frac{1}{N_1}\right), \tag{16}$$

where $f_C(E)$ is the final state distribution function for the correlated parking, since the normalized distribution function $f^*(E)$ remains exactly the same. The only effect of the added correlation is some increase of the number of particles in the final state that can be taken into account.

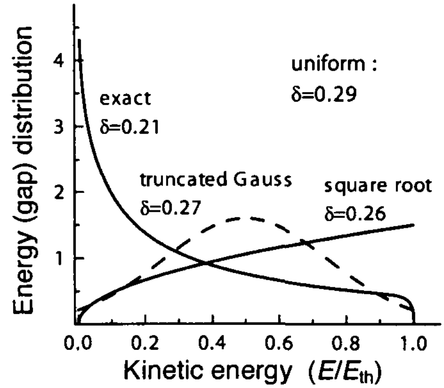


Figure 3. Comparison of the distribution functions used in different final state models together with corresponding mean-square deviations for the case of a narrow valence band.

Figure 3 compares several model final-state distribution functions used in the literature.[13,14] Also shown is the mean square root deviation δ for each model. In terms of δ, the final-state model estimate of the Fano factor is $F = \delta^2$.

One can observe that even though all final-state models predict values of the Fano factor below 0.1, their predictive value is limited. Most significant is the fact that the exact one-particle distribution function gives the final-state (uncorrelated) value of the RPP Fano factor as $F_{\mathrm{unc}} = 0.044$, whereas the exact value of the Fano factor for RPP model is known[16] to be $F = 0.051$. Calculations based on the exact two-gap distribution function, taking into account correlations between *nearest-neighbor gaps only*, produce a correction to F_{unc} that brings F very close to the exact value.[16] This means that correlations between distant gaps are not as important as the nearest-neighbor correlations.

4. Two particle correlations and Fano factor energy dependence

As described above, the main drawback of the final-state models is the assumption of uncorrelated energy distribution between the secondary particles in the final state. To further elucidate this point, we return to the narrow valence band model represented by RPP. In this model, the residual energy of pair creation is distributed between two secondary electrons. Consider now a special case of random parking on a lot whose length is triple the size of a single car, see Fig 4. One can readily see that *two* cars will always park in this lot, with no fluctuation of this number (the unique case of three tightly parked cars with no gaps has zero probability). Accordingly, when in a random sequence of energy branching a particle of energy equal to $3E_{\mathrm{th}}$ is created, the next energy branching will produce exactly 2 particles with no fluctuation, irrespectively of the fluctuating kinetic energies of these particles. When the initial energy is close to $3E_{\mathrm{th}}$, the two-particle final state will dominate. Correlations of this type will have an influence on the fluctuations of the final number of particles in all cases when at some intermediate stage one of the created secondaries has a small energy. This clearly compromises the validity of Eq. (3). In principle, one can take such correlations into account in the kinetic approach. One needs to study the kinetics of the fluctuations of the distribution function; in such a way that the source of these fluctuations is itself modified by the correlations in the energy distribution.[21] This approach is mathematically involved and has not yet been fully developed.

Figure 4. Random parking on a line $L = 3$. Two cars are always able to park, while three cars can never fit. Therefore, even though the spacing between cars fluctuates, there are no fluctuations of the number of cars.

Next, let us consider the *recursive* approach, another purely statistical technique specifically developed for branching processes. It studies the moments of the final-state energy distribution as functions of the initial energy. Due to the self-averaging properties of the branching process, the moments of distribution for the state emerging from high values of the initial particle energy can be recursively evaluated through the moments of the distribution for smaller initial energy. The recursive equations for the RPP were first derived by Rényi[22] and applied to the energy branching problem by van Roosbroek.[17]

The structure of recursive equations can be nicely illustrated in the RPP model. Consider the mean square of the number of cars placed along a line of length L. If there is a car at a distance x from the left end of the line, then the whole line is divided by this car into two lines of lengths x and $(L - x - 1)$. The average $<N^2(L)>$ can then be written in the form

$$\left\langle N^2(L)\right\rangle=\left\langle\left[1+N(x)+N(L-x-1)\right]^2\right\rangle. \tag{17}$$

Here the average is taken not only over all branching realizations for the intervals x and $(L - x - 1)$, but also over the variation of x in the range $0 < x < (L - 1)$. Since both the left and the right sides of this division are equivalent, we find

$$\left\langle N^2(L)\right\rangle=1+2\left\langle N(x)^2\right\rangle+4\left\langle N(x)\right\rangle+2\left\langle N(x)N(L-x-1)\right\rangle. \tag{18}$$

The last term in Eq. (18) correlates the number of cars that will be subsequently placed in the intervals x and $(L - x - 1)$. This correlation is negligible for large L and x, but important for small L and x and thus affects the resulting value of the Fano factor. Equation (18) should be accompanied by an equation for the mean $<N(L)>$, which by the same reasoning is of the form

$$\left\langle N(L)\right\rangle=\left\langle 1+N(x)+N(L-x-1)\right\rangle=1+2\left\langle N(x)\right\rangle. \tag{19}$$

Due to the uncorrelated nature of the branching itself, one can use the solution of Eq. (19) to calculate the correlation function in Eq. (18). We obtain

$$\left\langle \delta N^2(L)\right\rangle=2\left\langle N(x)^2\right\rangle+2\left\langle N(x)N(L-x-1)\right\rangle-4\left\langle N(x)\right\rangle^2, \tag{20}$$

Note that all terms in the right-hand side of Eq. (20) are of the order of N^2, but their difference is of the order of N. The correlation effects are therefore not negligible and sufficient accuracy is required in calculations. For the case of RPP and for several generalized RPP models, exact solutions of Eqs. (18–20) can be obtained[18,19,22] by using Laplace transformations of the recursive equations, which leads to differential equations for the Laplace transforms. These solutions are helpful in assessing the errors introduced by various approximations.

We have applied the statistical recursive approach to the calculation of the particle number fluctuations in the energy branching problem, considering the cases of a narrow valence band, symmetric (parabolic) bands, bands with constant DOS, and the case when the valence band has a finite energy.

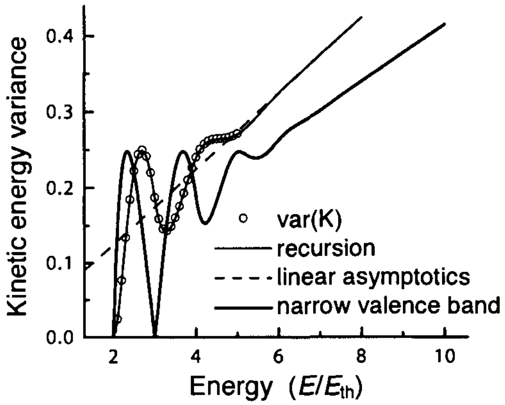


Figure 5. Dependence of the kinetic energy variance on the initial energy for a model with two symmetrical bands (dots on a line) and also for the case of a narrow valence band.

The kinetic energy variance is shown in Fig. 5 as function of the initial particle energy (which is needed for calculations of the Fano factor). For large $E_0 > 6E_{th}$ the variance grows linearly with E_0, corresponding to sub-Poissonian statistics with a constant Fano factor. For small $E_0 < 6E_{th}$, one see a strong variation of the variance indicating the importance of correlation effects. For the narrow valence band the variance goes to zero at $E_0 = 3E_{th}$, as discussed above. A dip in the variance near $3E_{th}$ remains also in the wide valence band model, although the effect there is less pronounced.

All exactly solvable models indicate the insensitivity of the results to the initial distribution for large energies, starting with $E > 5E_{th}$. This makes the recursive approach eminently tractable, even using realistic branching models. In the small range of energies where the results are sensitive to the initial distribution, one can obtain the solution of Eqs. (18–20) – with realistic models – by direct recursion, *i.e.* by starting from a small initial energy interval and then calculating the number of particles in a final state for progressively growing initial energy.

Calculations of the Fano factor are presented in Fig. 6. The variation $F(E_0)$ follows the kinetic energy variance. Values of the Fano factor vary in the range from $F = 0.051$ to $F = 0.763$ depending on the model and grow further when the threshold energy is shifted to higher energy due to the phonon emission.

The results of both the kinetic and the recursive approaches are supported by Monte Carlo modeling, in which one can exploit the fact (proven above) that in the final state the linear dependence on E_0 of both the average kinetic energy and its variance is exponentially accurate, starting from a reasonably small E_0, certainly not exceeding $E_0 \sim 10E_{th}$. As a result, both the average number of created pairs and its variance can be determined with Monte Carlo simulations with moderate values of E_0. To be sure, in order to achieve the same accuracy as that obtained for large E_0, the results must be averaged over a sufficient number of realizations. This, however, takes little memory or time.

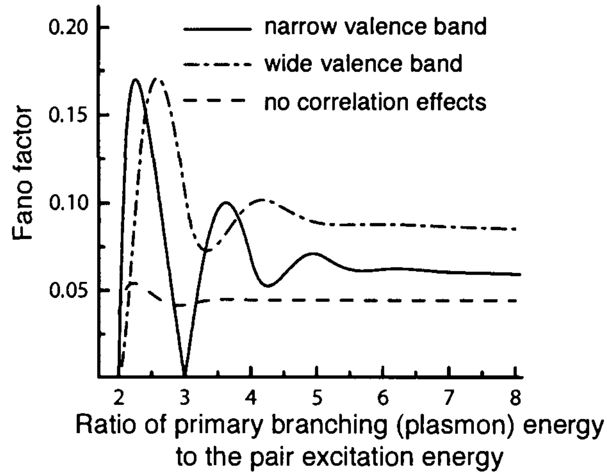


Figure 6. Fano factor dependence on the initial energy calculated recursively for a model with two symmetrical bands and for the case of a narrow valence band. Dotted line shows results obtained for the case of a narrow valence band in the final-state mode, using Eq. (5) with exact one-particle energy distribution function.

One can compare the computed values of the Fano factor with a hypothetical model in which the kinetic energy distribution of the particles in the final state is totally random within the $0 < E < E_{th}$ interval. Since this distribution retains constant total energy for any number of pairs it can be viewed as a distribution of hard rods of length E_{th} along a segment of fixed total length E_0. The random energy distribution should then be similar to the spatial distribution of particles shaped as hard rods in a one-dimensional gas (1D HR). For this problem one can calculate all multi-particle distribution functions,[23] which can be factored into products of pair correlation functions. The gaps in the 1D HR model are distributed in accordance with Poisson statistics, and F can be expressed in terms of the filling factor f (yield) exactly: $F = (1 - f)^2$. In this model, there is no jamming limit and the filling factor can be as large as $f = 1$, but Eq. (3) is exact. The 1D HR model can also be modified to allow for two-band branching and the phonon shift of the E_{th}.

The 1D HR model is an attractive alternative to the shot-glass model, since in this model the one-particle distribution function is flat, but the main effect of the Fano statistics – the distribution of a constant initial energy between energies of the particles in the final state – is already exactly incorporated. Since, in contrast to the shot-glass model, the 1D HR model requires no additional assumptions about the one-particle distribution, it can be viewed as a way to estimate the Fano factor for the known branching yield (the latter being determined by the pair excitation energy). For example, for a model of wide conduction and valence bands with square-root DOS, the yield is $f = 0.653$, giving $F = 0.120$. Inclusion of phonons further increases the Fano factor.

Consider the predictions of the 1D HR model for the exactly soluble RPP model. In this case, the jamming limit is exactly at $f = 0.74759...$ (the Rényi constant[23]) and the predicted Fano factor is $F = 0.064$. It is clearly an upper bound to the exact RPP value $F = 0.051$.

5. Discussion and conclusion

Our analysis shows that the range of variation of the Fano factor with parameters of the material is not very broad. One can point out the features that would lead to an anomalously high value of the Fano factor, above $F = 0.1$. It is more intriguing to find parameters that minimize the Fano factor. Our discussion suggests this may be achieved with a specially engineered alloy band structure, in which the valence band is either narrow or, which is more common, has sharp peaks in the density of states, so that the narrow valence band model becomes applicable. Correlation effects are more strongly pronounced in the narrow valence band limit. It would be most interesting to see some manifestation of the strong suppression of fluctuations by the correlation effects near $E_0 = 3E_{\rm th}$, illustrated in Fig. 4. One possibility would be to look for these effects in the dependence of noise in semiconductor x-ray detectors on the frequency $h\nu$ of incident x-ray flux of constant intensity. For $h\nu$ producing E_0 near $3E_{\rm th}$ one can expect suppression of the noise component associated with the branching of energy of photoabsorbed quanta.

There is also a tantalizing possibility to employ these correlations in practical γ-detectors, where the energy is, of course, much larger than $3E_{\rm th}$. The possibility relies on the established fact that at high electron energies the dominant loss mechanism – up to the final stage of the branching process – is plasmon emission. This brings the initial energy (for the final stage of energy branching through the impact ionization cascade) close to the plasmon energy which is ~16 eV in all common semiconductors. This energy is ~4.5 times larger than the pair excitation energy ε for Si, but only ~3.5 times larger than that for CdZnTe and only ~2.5 times larger than ε for GaP. Therefore, it is perhaps possible to find a compound crystal in which the ratio of the plasmon energy to $E_{\rm th}$ is optimal for the Fano factor reduction. The problem is not easy to handle theoretically with sufficient precision, because the threshold energy is based not only on the band structure but involves competition between phonon emission and impact ionization.

We conclude that correlation effects are important for the correct understanding of Fano statistics. Depending on the details of energy branching they can lead to either enhanced or suppressed value of the Fano factor and may be of value in the design of radiation detectors.

Acknowledgments

This work was supported by the Domestic Nuclear Detection Office (DNDO) of the Department of Homeland Security, by the Defense Threat Reduction Agency (DTRA) through its basic research program, and by the New York State Office of Science, Technology and Academic Research (NYSTAR) through the Center for Advanced Sensor Technology (Sensor CAT) at Stony Brook.

References

1. U. Fano, "Ionization yield of radiations II: The fluctuations of the number of ions," *Phys. Rev.* **72**, 26 (1947).
2. H. Spieler, *Semiconductor Detector Systems*, New York: Oxford University Press, 2005.
3. L. Rossi, P. Fisher, T. Rohe, and N. Wermes, *Pixel Detectors: From Fundamentals to Applications (Particle Acceleration and Detection)*, Berlin: Springer, 2006.
4. A. Kastalsky, S. Luryi, and B. Spivak, "Semiconductor high-energy radiation scintillation detector," *Nucl. Instr. Meth. Phys. Res. A* **565**, 650 (2006).
5. C. E. Lehner, Z. He, and F. Zhang, "4π Compton imaging using a 3-D position-sensitive CdZnTe detector via weighted list-mode maximum likelihood," *IEEE Trans. Nucl. Sci.* **51**, 1618 (2004).
6. S. Luryi, "Impregnated semiconductor scintillator", *Intern. J. High Speed Electronics Syst.* **18**, 973 (2008).
7. R. Devanathan, L. R. Corrales, F. Gao, and W. J. Weber, "Signal variance in γ-ray detectors – a review," *Nucl. Instr. Meth. Phys. Res. A* **565**, 637 (2006).
8. R. C. Alig, S. Bloom, and C. W. Struck, "Scattering by ionization and phonon emission in semiconductors," *Phys. Rev. B* **22**, 5565 (1980).
9. J. Sempau, J. M. Fernandez-Varea, E. Acosta, and F. Salvat, "Experimental benchmarks of the Monte Carlo code Penelope," *Nucl. Instr. Meth. Phys. Res. B* **207**, 107 (2003).
10. M. Vilches, S. Garcia-Pareja, R. Guerrero, M. Anguiano, and A. M. Lallena, "Monte Carlo simulation of the electron transport through thin slabs: A comparative study of PENELOPE, GEANT3, GEANT4, EGSnrc and MCNPX," *Nucl. Instr. Meth. Phys. Res. B* **254** 219 (2007).
11. H. Bichsel, "Straggling in thin silicon detectors," *Rev. Mod. Phys.* **60**, 663 (1988).
12. D. V. Jordan, A. S. Reinolds, J. E. Jaffe, K. K. Anderson, L. R. Corralesa, and A. J. Peurrung, "Simple classical model for Fano statistics in radiation detectors," *Nucl. Instr. Meth. Phys. Res. A* **385**, 146 (2008).
13. C. Klein, "Bandgap dependence and related features of radiation ionization energies in semiconductors," *J. Appl. Phys.* **39**, 2029 (1968).
14. H. R. Bilger, "Fano factor in Ge at 77 K," *Phys. Rev.* **163**, 238 (1967).

15. W. E. Drummond and J. L Moll, "Hot carriers in Si and Ge Detectors," *J. Appl. Phys.* **42,** 5556 (1971).
16. A. Subashiev and S. Luryi, "Correlation effects in sequential energy branching: An exact model of the Fano statistics," arXiv:0911.1532vl, submitted to *Phys. Rev. E* (2009).
17. W. van Roosbroeck, "Theory of the yield and Fano factor of electron-hole pairs generated in semiconductors by high-energy particles," *Phys. Rev.* **139,** A1702 (1965).
18. A. V. Subashiev and S. Luryi, "Random sequential adsorption of shrinking or expanding particles," *Phys. Rev. E* **75**, 011123 (2007).
19. A. V. Subashiev and S. Luryi, "Fluctuations of the partial filling factors in competitive RSA from binary mixtures," *Phys. Rev. E* **76,** 011128 (2007).
20. S. Picozzi, R. Asahi, C. B. Geller, A. Continenza, and A. J. Freeman, "Impact ionization in GaAs: A screened exchange density-functional approach", *Phys. Rev. B* **65**, 113206 (2002).
21. Sh. Kogan, *Electronic Noise and Fluctuations in Solids*, Cambridge, UK: Cambridge University Press, 1996.
22. A. Rényi, "On a one-dimensional problem concerning random space-filling," *Publ. Math. Inst. Hung. Acad. Sci.* **3**, 109 (1958).
23. Z. W. Salsburg, R. W. Zwanzig, and J. G. Kirkwood, "Molecular distribution functions in a one-dimensional fluid," *J. Chem. Phys.* **21**, 1098 (1953).

This page intentionally left blank

The Future of Microelectronics is ... Macroelectronics

M. A. Alam, N. Pimparkar, and B. Ray
Dept. of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47906-3800, U.S.A.

1. Introduction

The title is actually not as provocative as it might seem. No one debates the miraculous success of microelectronics – enshrined in Moore's law and iconified by Pentium chips – in transforming how we compute and communicate. Even as the technology – buffeted by many fundamental limits – slides into a more mature phase, the degree of miniaturization of transistors continues to amaze, *e.g.*, the number of transistors made in a year by Intel or TSMC exceeds the estimates of ants or leaves in the world! Yet, the most exciting trends in electronics today are defined not by computers, but by flat panel displays, solar cells, and bioelectronics sensors. Sizing the solar cells, batteries or flat panel TV to the dimension of a Pentium is obviously counterproductive, if fact for such systems – the bigger, the better. If we categorize the electronics of flat-panel displays, energy conversion and conservation devices as being macro- or large-scale electronics[1] (see Fig. 1), presumably few would quibble with the title of the article that "the future of microelectronics lies with macroelectronics".[2,3]

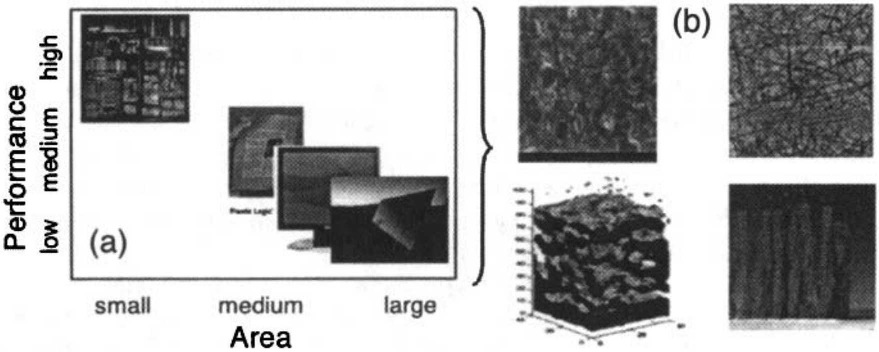


Figure 1. (a) Traditional microelectronic ICs emphasize high performance with small footprint (top, left), but macroelectronic devices like displays and energy conversion devices favor large area devices with low to moderate performance (bottom, right). (b) Material innovations, clockwise from top left: polysilicon transistors for displays,[1] carbon nanonets for TFTs,[2,4] mesoporous electrodes for super-capacitors,[5] and donor-acceptor blends for solar cells[8] have defined the rise of macroelectronic systems.

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

While the emergence of macroelectronics as the new economic driver for the industry is plain to see, few appreciate the fundamental conceptual/design challenges such electronic components engender. Indeed, broad/economic adoption of macroelectronics would rely on development of theory, process, and device concepts no less transformative than those of microelectronics industry. The key distinguishing feature of macroelectronics is that the economics of material/energy dictate that macroelectronics devices use thin films processed at the lowest possible temperature.[1,2] On the theory side, it means letting go of the presumptions of long-range periodicity supported via Bloch's theorem, Ohm's law, and isotropic transport, *etc.* that have been so central to the analysis of single crystalline microelectronic devices. Rather, one needs to view the optimization of macroelectronic devices made of amorphous/roughened/gyroid materials (solar cells and thermoelectrics) and mesoporous structures (batteries/super-capacitors) in a fundamentally new way. *In sum, materials for macroelectronics should not be considered as a poorer cousin of the single crystalline silicon, but rather a new system with unique properties that can be used to broaden the reach of electronics in everyday life.*

In this chapter, we consider two illustrative examples to demonstrate the nontrivial theoretical framework required for the analysis and design of macroelectronic devices. In Section 2, we consider a thin-film transistor (TFT) technology based on network of carbon nanotubes (CNTs) with potential applications in flexible electronics.[2,6,7] This example will illustrate why macroscopic Ohm's law does not apply to macroelectronic materials and demonstrate the power of nonlinear percolation theory in explaining the transport properties of the TFTs based on the nanonet technology. In Section 3, we discuss the process and device models for bulk-heterostructure solar cells,[9] because process induced non-classical morphology[10] dictates the device efficiency of these solar cells. And we will also show how a counterintuitive transformation makes the nonclassical transport problem amenable to classical analysis. We conclude in Section 4 by suggesting a deep analogy between carrier transport in macroelectronic devices and the nutrient transport in biological entities and explain how the requirement that they both operate close to thermodynamic equilibrium forces them to adopt unconventional geometrical solutions to the diffusive transport problems. For a broader conceptual framework and more detailed analysis of this class of problems, the readers may wish to consult 2009 Intel/Purdue Summer Lecture Series entitled "Randomness and variability in electronic systems".[11]

2. Fabrication and physics of nanonet transistors

The channel of macroelectronic TFTs (see Fig. 2) are generally made of CVD-deposited a-Si,[12] solution processed organics,[13,14] transfer-printed crystalline semiconductors,[15,16] *etc.* More recently, a new material based on a network of CNTs (nanonet TFTs) has emerged as higher performance alternative to traditional options.[4,7,17] To develop these films, high-mobility, nearly crystalline CNTs are grown either by vapor-liquid-solid method or by laser arc discharge, then purified by oxidation, dissolved in solvent, debundled by sonication and surfactants, and finally solution-cast at low temperature onto plastic or glass substrates, see Fig. 2. The decoupling of crystal growth (at high temperature) and film deposition (at low temperature) makes high-performance TFTs possible. The transistors may be gated either from the top or the back, typically via Pd and Au as source/drain and gate contacts, respectively.

In achieving this high performance, however, the nanonet TFTs highlight the consequences of the randomness and variability characteristic of all TFT channel materials. For example, unlike classical Ohmic transistors, (i) the channel resistance does not scale linearly with width W,[18] nor does it scale inversely with channel length L;[19,20] (ii) highly aligned tubes carry less current than randomly oriented tubes;[21] (iii) striping the channel (without changing the width) changes the current dramatically,[22] *etc.* These features reflect the characteristic nonhomogeneous morphology of the nanonet-TFTs that cannot be understood by classical transistor theory[30] developed for isotropic materials like crystalline silicon.

Instead of reducing the nanonet channel to a homogenized effective media defined by classical transport parameters, one may consider the morphology of the network as an irreducible feature of the transport problem and explore the electron transport through nanonets via stick-percolation.[11,23-25] In this approach, electrons would cross the channel by diffusing through a series of interconnected sticks (Fig. 2, top). The theories of percolation and stochastic geometry may appear to be ideally suited to treat this problem – however, several issues prevent straight-

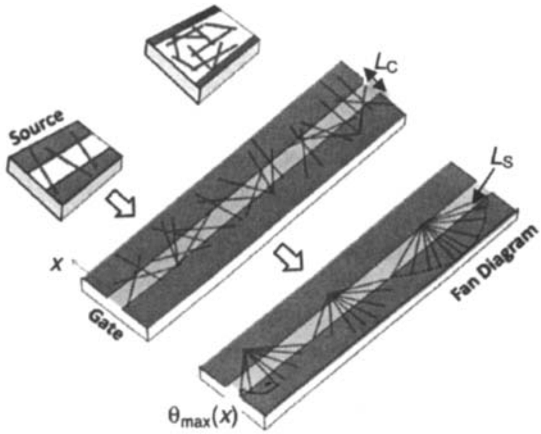


Figure 2. In short channel TFTs (left), a single nanotube or nanowire (stick) can directly bridge the channel, but not in long channel TFTs (top). The short channel TFTs have randomly oriented sticks germinating at different locations x from the channel/source edge (middle figure). The tubes can be rearranged (without changing the germination distance from channel edge and its orientation) to create a series of "fan diagrams". The maximum angle of each fan (θ_{max}) depends on L_S, L_C, and x, as shown in the bottom figure.

forward application of the classical percolation models: classical percolation has been developed for infinite systems,[26] while electron transport in nanonet transistors involve finite systems; the classical theory presumes perfect, zero resistance transfer between one stick to next, while in practice tube-tube resistance may dominate transport;[21] finally, and most importantly, classical percolation is a linear response theory with the *presumption* of small drain biases and sticks being defined by a constant conductance, so it does not reflect the reality of large gate/drain biases typical of transistor operation and the invasion of a pinch-off region (and non-uniform conductance) being an important part of the transport problem. No wonder that percolation theory has never been used in describing transport in TFTs.

One can make progress in developing a percolation theory of modern transistors by considering ballistic transport in *short* channel transistors where randomly oriented tubes of density D_C and length L_S directly bridge the channel of length L_C.[27,28] As shown in Fig. 2, the random orientation of the tubes allows one to (conceptually) collect the sticks in a series of fan diagrams, with germination points fixed distance x away from the source/gate edge. Each of these "fans" makes a maximum bridging angle of $\theta_{max}(x) = \cos^{-1}(L_C + x/L_S)$, which can be summed in a manner similar to the Buffon needle problem:[29]

$$N_S = \frac{2}{\pi} \int_0^{L_S - L_C} D_C \, dx \, \theta_{max}(x) = \frac{2 D_C L_S}{\pi} \left[\sqrt{1 - R^2} - R \cos^{-1} R \right], \qquad (1)$$

with $R \equiv L_C/L_S$. If we further assume that $L_S \leq \lambda$ – the mean free path – the transport can be presumed ballistic and the drain current $I_D^{(n)}$ in the nth tube can be calculated from the Landauer formula:[30]

$$I_D^{(n)} = \frac{q}{\pi\hbar} \int_{E_B(V_G)}^{\infty} dE \, [f_0(E_{FS}) - f_0(E_{FD})] \quad \equiv \quad \varphi(V_G, V_D)$$

where f_0 is the Fermi-Dirac function, while E_{FS} and E_{FD} are the source and drain Fermi levels, respectively. The total drain current is obtained by summing over the "fans", *i.e.*,

$$I_D = \int_0^{L_S - L_C} \frac{2 D_C \, dx}{\pi} \theta_{max} f = \varphi(V_G, V_D) \times \xi(D_C L_S^2, L_C/L_S), \qquad (2)$$

where

$$\xi \equiv \frac{2 D_C L_S}{\pi} \left[\sqrt{1 - R^2} - R \cos^{-1} R \right].$$

Equation (2) generalizes the classical percolation theory to the nonlinear regime relevant for transistors. Focusing on the geometrical factor ξ, one finds that electron transport in these systems would depend on two scaling variables: normalized stick density $D_C L_S^2$ and the ratio $R = L_C/L_S$ of the stick length to the channel length.

At the first glance, Eq. (2) appears reasonable because the drain current should indeed increase if the either the density or the length of tubes is increased with respect to the channel length. A closer examination of Eq. (2), however, suggests that the result actually is far from trivial: when plotted as a function of L_C, the drain current I_D scales as $\sim 1/L_C$ despite the fact that in deriving Eq. (2), electron transport was *presumed ballistic*! This anomalous channel length scaling reflects a deep and fundamental characteristic of percolating network: the drain current scales with $1/L_C$ not because there is additional scattering in longer channel transistors, making the transport Ohmic, but because a number of sticks (N_S) drops out of the "fan" in Fig. 2 because they can no longer bridge the S/D contact distance in a longer channel TFT. For ballistic transport $\xi = N_S$, so the reduction in the number of sticks available for conduction is directly reflected in the drain current.

Even though Eq. (2) was derived from considering short-channel ballistic transistors, we find that this remarkable result survives scrutiny by numerical simulation and experimental results. Indeed, the scaling of N_S holds even if the transport in the individual sticks is diffusive.[28] Here however the mobility itself depends on the channel length. Taken together, I_D decreases super-linearly with L_C in diffusive nanonet transistors. Finally, in an even longer channel transistors (see Fig. 2, top), where no single tube can bridge the channel, but rather electrons must percolate through a network of sticks to go from source to drain, numerical simulation shows that

$$\xi \sim \left(\frac{L_S}{L_C}\right)^{m(D_C L_S^2)} / L_S \,. \tag{3}$$

The appearance of the same two scaling variables in Eq. (3) is reassuring and the value of the exponent m is ~ 1 at high stick density (Fig. 3, circles) and ~ 2 for density close to the percolation threshold (Fig. 3, squares). The classical result $\xi = 1/L_C$ is easily reproduced in the high-density $m \sim 1$ limit.[21] Moreover, this approach explains the puzzlingly large spread in mobility values reported in TFTs. Classical mobility formulae are based on the assumption of Ohmic conduction in thin films ($m \sim 1$) *regardless of the stick density*, a condition that is not satisfied in nanonet transistors. The artifact of the mobility spread disappears within the theoretical framework of nonlinear percolation theory.[21]

Apart from the nontrivial channel length scaling, the most remarkable aspect of Eq. (2) is that the total current is simply a *product* of a geometry factor (stick length, channel length, stick density, *etc.*) and a voltage-dependent transport term, see Fig. 3. It means that if we find the voltage function from classical transport theory (*e.g.* Shockley's standard transistor theory[31]) and the geometry factor from linear percolation theory,[26] we should be able to describe TFTs with percolating channels at arbitrary bias conditions.

We conclude this section by emphasizing that transport in nanonet transistors built from CNTs involves heterogeneous percolation: two-thirds of the tubes are semiconducting (can be turned on and off by the gate, see Fig. 2) and

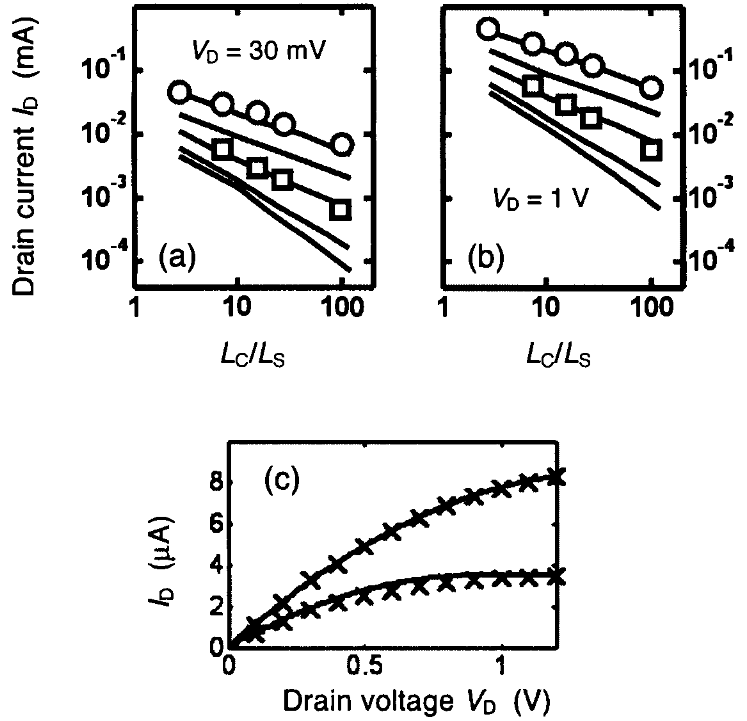


Figure 3. The geometric component ξ of Eq. (2) is predicted to be independent of transistor biasing, as shown both numerically (lines) and experimentally (high density: circles, low density: squares) for a wide variety of L_C, for V_D = 30 mV (a) and V_D = 1 V (b). Likewise, for two different device geometries, while ξ of Eq. (2) is expected to be different, the shape of the voltage scaling factor φ is expected to be the same (c): two devices with different densities and L_C are shown, lines are theory, symbols are experimental.

one-third metallic (cannot be gated off). At very high densities, both subnetworks percolate and the metallic subnetwork shorts the TFT, making the on/off ratio unacceptably low; at very low densities, neither of subnetworks percolates, so that the drain current is zero. The transistors must be designed at optimum intermediate density so that the semiconducting subnetwork would percolate, but the metallic tubes would not. Optimizing the density of such heterogeneous network beyond the reach of classical percolation theory, especially for integrated circuits with many transistors. The nonlinear percolation theory discussed above offers an innovative route to controlling percolation threshold (see Fig. 4) by "striping" the channel in such a way that the metallic subnetwork is disrupted without corresponding loss in current in the semiconducting subnetwork.[2] Powerful concepts like "striping" arises only if the CNT network is viewed for what it is – a perfect percolating network, not an imperfect single crystalline material. For further details, we refer the reader to the lecture series posted at Ref. 11.

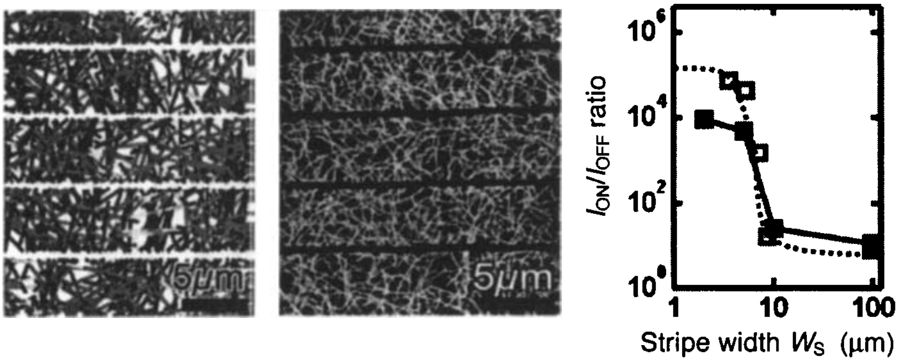


Figure 4. Both simulation (left) and experiment (middle) show that if a percolating network (top view) of metallic (blue) and semiconducting (red) tubes is striped by a laser (horizontal lines), resulting in stripe width W_S, the changes in the percolation thresholds of the respective subnetworks dramatically increase the on/off ratio of the transistors (right figure: black line – experiment, open squares – theory).[2]

3. Bulk heterostructure solar cells

As the second example, we consider bulk heterostructure (BH) solar cells[8] to demonstrate the importance of morphology in macroelectronic devices. Indeed, BH provides a perfect illustration of the famous biological dictum that "form defines function". As shown in Fig. 5, classical solar cells are made of simple, single crystalline *pn* or *p-i-n* parallel junctions in which the built-in field separates the electron-hole pairs (generated by incident sunlight) to drive an output load.[32] Currently, the solar cell market is dominated by such cells.[12,33] However, the efficiency limit of ~30%, the cost of the single crystalline silicon, the need for high temperature processing, and most importantly, the cost of installation of bulky panels, etc. make the cells more expensive than competing sources of electricity (solar cells are still to reach "grid-parity"). A BH solar cell is an intimately mixed dual-polymer structure (typically PCBM and P3HT) that offers a lower-cost alternative. The lower efficiency of polymer cells is potentially counterbalanced by low-cost solution processing and ease of installation of lightweight panels at relatively low cost.

In BH cells, sunlight creates electron-hole pairs in a manner similar to the crystalline cells. However, unlike 3D crystals, the quasi-1D polymers do not dilute the Coulomb attraction between the newly created electrons and holes; therefore the two particles diffuse as a pair (an exciton). Unless the pair is rapidly separated via the strong electric field of a heterojunction between donor/acceptor polymers, the excitons will be lost to recombination. Parallel junctions similar to crystalline cells are ineffective, because only the excitons created within the junction diffusion length (20–30 nm) can be converted to useful photocurrent.

Unlike crystals, however, polymers are solution processed and malleable. Once the two polymers are homogenized and annealed to phase-segregation, the surfaces will lie within a tens of nm from each other, and hence a majority of the excitons can in principle be collected and most photons would contribute to photocurrent.[34] This elegant idea of morphology-engineering lies at the heart of a decade-long effort in BH solar cell research. Indeed, morphology engineering plays the same role for macroelectronic devices as doping engineering does for classical electronics in shepherding the charge carriers to the respective contacts.

Unfortunately, the same intertwined morphology that promotes exciton collection also makes it difficult for the separated particles to traverse the maze to reach the electrodes (see Fig. 5). Moreover, unconnected or wrongly connected islands are self-excluded via build-up of trapped charges, again see Fig. 5, right. Optimized morphology must balance the needs for exciton dissociation and charge collection; so far, the optimization has been empirical and difficult.[35]

The regularity of phase segregation of well-blended P3HT/PCBM polymers and the conservation of phases suggest that the process may (at least approximately) be described by spinodal decomposition, a phenomenological description of which is given by the Cahn-Hilliard equation:[36]

$$\frac{\partial \vartheta}{\partial t} = M_0\left(\nabla^2 \frac{\partial f}{\partial \vartheta} - 2\kappa\nabla^4\vartheta\right),$$

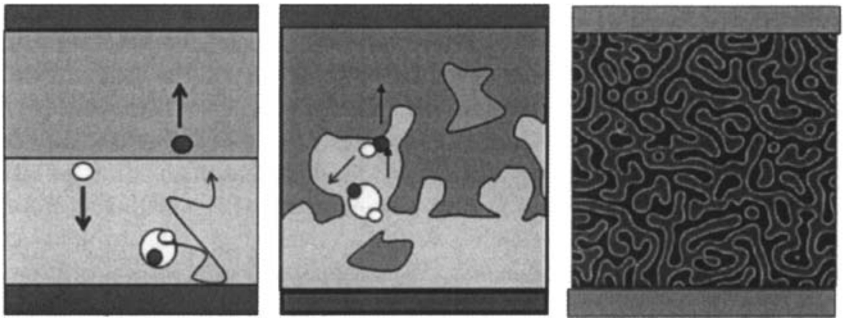


Figure 5. Classical solar cells operate by generating electron-hole (e-h) pairs (open and filled circles) and separating the carriers through a built-in field of a planar junction (left). If the e-h pairs are strongly bound by Coulomb interaction, they may recombine before separation; a well-mixed nonplanar morphology makes dissociation interfaces accessible within the exciton diffusion length and improves charge collection (middle); numerical simulation of the phase-segregation process via the Cahn-Hilliard equation demonstrates the existence of well-mixed morphology helpful for exciton dissociation (right). However, the difficulty for electrons and holes to reach respective contacts at the top and bottom, exclusively via the dark and light channels is also obvious and poses an optimization challenge.

where $\vartheta(x,t)$ is the volume fraction of one of the polymers, M_0 is the Onsager mobility coefficient dictated by the polymers and the solvent, and κ is the gradient energy coefficient dictated by interfacial energy between two phases. The free energy function f of the homogenized mixture depends on entropy and interaction energy of the solution,[10] as well as the relative sizes of the polymer chains. Except for the initial transient phase which is fractal,[37] the size evolution is characterized by a length scale W_C determined by the Lifshitz-Slyozov law:[38]

$$W_C \sim [M_0 f'' \sqrt{\kappa \Delta f}\, t_A]^n,$$

where t_A is the anneal time, Δf is the change in free energy, and f'' is the second derivative of the free energy with respect to the volume fraction. Given that the phase volume is conserved, the effective surface area S scales as $W_C(t)S(t)$ = a constant, so $S(t) \sim (W_C)^{-1} \sim t_A^{-n}$. The surface area reaches a maximum in the well-mixed film, but it reduces as annealing coarsens the film.

It is easy to calculate the efficiency I_{EX} of the exciton dissociation at the donor/acceptor interface as $I_{EX} \sim G_{EX}L_{EX}/W_C \sim t_A^{-n}$. In other words, the efficiency of exciton dissociation reduces with anneal time, an intuitive result. Both experimental and numerical simulations confirm this prediction. This result may suggest that it is counterproductive to anneal the sample, because finer interfaces lead to better exciton dissociation. *This is unfortunately not true.* One can show via a one-dimensional analysis of the Cahn-Hilliard equation that there is a minimum time to sufficiently phase-separate the interface boundaries to produce sharp heterobarrier discontinuities. Since exciton binding energy must be smaller than the heterointerface discontinuity (otherwise exciton dissociation is suppressed exponentially[39]), a minimum annealing time is required for boundary formation and efficient phase separation.

For a very thin, 2D lateral BH cells, the interface formation time and the phase segregation time define the optimum annealing time. For typical 3D films, however, there is the additional constraint that the *charge collection following exciton dissociation must be efficient as well.* Here fine-grained structure, optimal for exciton dissociation, leads to many blind allays for charge collection (Fig. 5, right panel). Only after the domain size W_C becomes comparable to film thickness will a percolating path offer easy passage to the electrodes for a substantial fraction of charge carriers. Numerical calculations show[9] and empirical evidence suggests[35] that this optimum annealing time is approximately 10–15 minutes for 1:1 polymer blends at typical solvent concentration. Indeed, this modeling and simulation allows us view the mysterious process of BH formation as a predictable consequence of factors like the polymer volume fraction, solvent concentration, anneal temperature, energy discontinuity at the polymer boundary, *etc.* The fundamental role of morphology in these films is obvious and has no analog in transport literature of classical devices.

One final point. Although both the nanonet transistors and BH solar cells appear to be random percolating structures, there are different in two important aspects: the morphology in nanonet transistors depends on the density of sticks, but this spatial distribution is time invariant and fractal *at all stick densities.* This scale-invariant morphology is the cornerstone of percolation analysis of nanonet TFTs. For BH solar cells, on the other hand, the relative concentration of the two phases is time-variant, but the morphology is not. And after the initial transient of fractal morphology,[36] *the structure is dominated by a single length scale.* Classical percolation theory has little to say about transport in such structures.

5. Conclusions

In this chapter, we have discussed the growing importance of macroelectronic devices for the semiconductor industry. Macroelectronics use materials processed at lower temperature which has led to the historical perception that macroelectronic materials can be viewed as an inferior version of the microelectronic materials and that their transport properties can be treated within the same framework of microelectronic devices, albeit with lower mobility and higher trap density. On the contrary, in this paper we have shown that the malleability of many macroelectronic materials offers new opportunities of device optimization where one can improve the intrinsic material performance through application-specific tailoring of the morphology/geometry of the material. *In this sense, macroelectronic material is biomimetic*: Nature functions close to thermodynamic equilibrium and the machines She creates (*e.g.*, jellyfish) use the geometry of branching to maximize the diffusive capture of nutrient, while balancing it against the energy penalty to transport food through an over-branched structure. This unique/active role of geometry in pervasive in Nature, and similar active role of geometry in device design distinguishes macroelectronics from the classical microelectronic counterparts and offers opportunities for innovative/imaginative optimization beyond the reach of classical electronics.

Acknowledgments

This work was supported in part by the Network of Computational Nanotechnology and Columbia NSEC programs. We also acknowledge a long and very enjoyable collaboration with Prof. John Rogers at Univ. of Illinois–Urbana-Champaign on nanonet transistors.

References

1. T. I. Kamins, *Polycrystalline silicon for integrated circuits and displays*, 1998.
2. Q. Cao, H. S. Kim, N. Pimparkar, *et al.*, "Medium scale carbon nanotube thin film integrated circuits on flexible plastic substrates," *Nature* **454**, 495 (2008).
3. R. H. Reuss, D. G. Hopper, and J. G. Park, "Macroelectronics," *MRS Bull.* **31**, 447 (2006).

4. S.-H. Hur, C. Kocabas, A. Gaur *et al.*, "Printed thin-film transistors and complementary logic gates that use polymer-coated single-walled carbon nanotube networks," *J. Appl. Phys.* **98**, 114302 (2005).
5. C. Du and N. Pan, "High power density supercapacitor electrodes of carbon nanotube films by electrophoretic deposition," *Nanotechnology* **17**, 5314 (2006).
6. L. Hu, D. S. Hecht, and G. Gruner, "Percolation in transparent and conducting carbon nanotube networks," *Nano Lett.* **4**, 2513 (2004).
7. G. Gruner, "Carbon nanotube films for transparent and plastic electronics," *J. Mater. Chem.* **16**, 3533 (2006).
8. P. Peumans, S. Uchida, and S. Forrest, "Efficient bulk heterojunction photovoltaic cells using small-molecular-weight organic thin films," *Nature* **425**, 158 (2003).
9. S. R. Forrest, "The limits to organic photovoltaic cell efficiency," *MRS Bull.* **30**, 28 (2005).
10. B. Ray, P. R. Nair, R. E. García, *et al.*, "Modeling and optimization of polymer based bulk heterojunction (BH) solar cell," to appear in *Tech. Dig. IEDM* (2009).
11. M. A. Alam, "Nanostructured electronic devices: Percolation and reliability," available at *http://nanohub.org/resources/7168* (2009).
12. P. G. Borden, "Laser applications in photovoltaics," *Photovoltaics World* (2009).
13. A. Dodabalapur, L. Torsi, and H. Katz, "Organic transistors – 2-dimensional transport and improved electrical characteristics," *Science* **268**, 270 (1995).
14. J. Schon, A. Dodabalapur, Z. Bao, *et al.*, "Gate-induced superconductivity in a solution-processed organic polymer film," *Nature* **422**, 92 (2003) [retraction of *Nature* **410**, 189 (2001)]."
15. J. Yoon, A. Baca, S. Park, *et al.*, "Ultrathin silicon solar microcells for semitransparent, mechanically flexible and microconcentrator module designs," *Nature Mater.* **7**, 907 (2008).
16. M. Meitl, Z. Zhu, V. Kumar, *et al.*, "Transfer printing by kinetic control of adhesion to an elastomeric stamp," *Nature Mater.* **5**, 33 (2006).
17. E. Artukovic, M. Kaempgen, D. S. Hecht, *et al.*, "Transparent and flexible carbon nanotube transistors," *Nano Lett.* **5**, 757 (2005).
18. C. Kocabas, N. Pimparkar, O. Yesilyurt, *et al.*, "Experimental and theoretical studies of transport through large scale, partially aligned arrays of single walled carbon nanotubes in thin film type transistors," *Nano Lett.* **7**, 1195 (2007).
19. S. Kumar, J. Y. Murthy, and M. A. Alam, "Percolating conduction in finite nanotube networks," *Phys. Rev. Lett.* **95**, 066802 (2005).
20. E. S. Snow, J. P. Novak, P. M. Campbell, *et al.*, "Random networks of carbon nanotubes as an electronic material," *Appl. Phys. Lett.* **82**, 2145 (2003).
21. N. Pimparkar, S. Kumar, Q. Cao, *et al.*, "Current-voltage characteristics of long-channel nanobundle thin-film transistors: A 'bottom-up' perspective," *IEEE Electron Dev. Lett.* **28**, 157 (2007).

22. S. J. Kang, C. Kocabas, T. Ozel, *et al.*, "High performance electronics based on dense, perfectly aligned arrays of single walled carbon nanotubes," *Nature Nanotechnol.* **2**, 230 (2007).
23. M. A. Alam, "The long and short of pick-up stick transistors: A promising technology for nano- and macroelectronics," available online at *http://nanohub.org/resources/1214*.
24. A. Behnam, L. Noriega, Y. Choi, *et al.*, "Resistivity scaling in single-walled carbon nanotube films patterned to submicron dimensions," *Appl. Phys. Lett.* **89**, 093107 (2006).
25. M. Topinka, M. Rowell, D. Goldhaber-Gordon, *et al.*, "Charge transport in interpenetrating networks of semiconducting and metallic carbon nanotubes," *Nano Lett.* **9**, 1866 (2009).
26. D. Stauffer, and A. Aharony, *Introduction to Percolation Theory*, London: Taylor and Francis, 1992.
27. M. A. Alam, N. Pimparkar, S. Kumar, *et al.*, "Theory of nanocomposite network transistors for macroelectronics applications," *MRS Bull.* **31**, 466 (2006).
28. N. Pimparkar, J. Guo, and M. A. Alam, "Performance assessment of sub-percolating nanobundle network transistors by an analytical model," *IEEE Trans. Electron Dev.* **54**, 637 (2007).
29. A. M. Mathai, *An Introduction to Geometrical Probability: Distributional Aspects with Applications*, Amsterdam: Gordon & Breach, 1999.
30. S. Datta, *Quantum Transport: Atom to Transistor*, 2nd ed., New York: Cambridge University Press, 2005.
31. R. F. Pierret, *Semiconductor Device Fundamentals*, New York: Addison-Wesley, 1996.
32. P. Bhattacharya, *Semiconductor Optoelectronic Devices*, 2nd ed., Upper Saddle River, NJ: Prentice Hall, 1997.
33. M. Green, "Crystalline and thin-film silicon solar cells: State of the art and future potential," *Solar Energy* **74**, 181 (2003).
34. S. Park, A. Roy, S. Beaupre, *et al.*, "Bulk heterojunction solar cells with internal quantum efficiency approaching 100%," *Nature Photonics* **3**, 297 (2009).
35. H. Hoppe and N. Sariciftci, "Morphology of polymer/fullerene bulk heterojunction solar cells," *J. Mater. Chem.* **16**, 45 (2006).
36. R. W. Balluffi, S. M. Allen, W. C. Carter, *et al.*, *Kinetics of Materials*, Hoboken, NJ: Wiley-Interscience, 2005.
37. W. Klein, "Fractals and multifractals in early-stage spinodal decomposition and continuous ordering," *Phys. Rev. Lett.* **65**, 1462 (1990).
38. I. Lifshitz and V. Slyozov, "The kinetics of precipitation from supersaturated solid solutions," *J. Phys. Chem. Solids* **19**, 35 (1961).
39. L. Onsager, "Initial recombination of ions," *Phys. Rev.* **54**, 554 (1938).

An Integration Challenge: Information and Communication Technologies to Address Indoor Air Quality in Commercial Buildings

M. D'Iorio
National Research Council Canada, Institute for Microstructural Sciences
Ottawa, Ontario, Canada, K1A 0R6

1. Introduction

Much attention has been drawn to health problems related to airtight building environments with poor indoor air quality or limited ventilation capabilities.[1] This problem is encountered not only in commercial buildings, but also in homes where modern construction materials have exacerbated the problem. In the last century, the air tightness of a typical Canadian building has improved by 48% between 1900 and 1960 and a further 25% between 1960 and 2004, driven in part, by the energy crisis of the 1970's. Figure 1 illustrates this trend for three Canadian cities.[2]

Indoor air pollutants include tobacco smoke, dust mites, animal dander, molds, combustion gases, volatile organic compounds (VOC) and heavy metal dust. The building and insulating materials that allow us greater comfort in our homes also outgas contaminants that are deleterious to our health. Some of the sources of contamination are shown in Table 1.

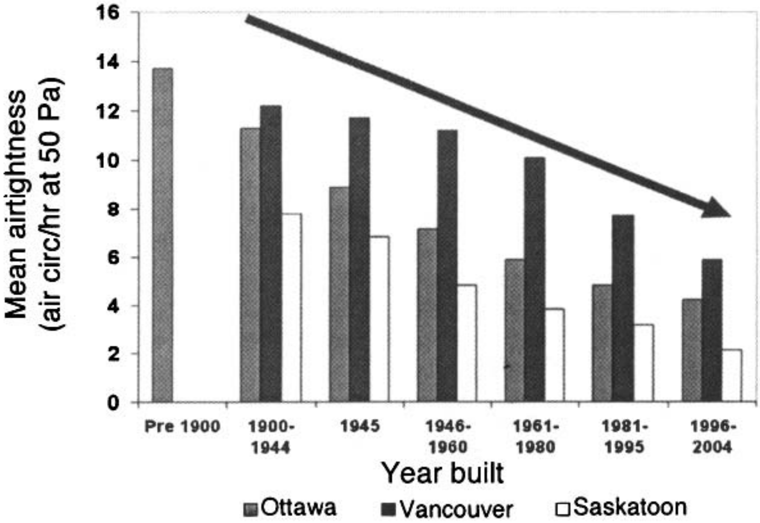


Figure 1. Mean air tightness in buildings in three Canadian cities measured in air circulation per hour at a pressure of 50 Pascals.[2]

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

Construction materials, cleaning products, office equipment	**Human beings/ activities, animals & plants**	**Outdoor air pollution**
• Formaldehyde • Volatile organic compounds (VOCs) e.g. aromatic compounds • Particles (laser printers) • Insecticides • Pesticides	• CO_2 • Odors • Bacteria and viruses • VOCs (cosmetics) • Odors (waste bins etc.) • Moulds • Spores • Skin particles • Smoking	• Ozone, photochemical smog • Particles (traffic) • Carbon monoxide • Carbon dioxide • NO/NO_2 (NO_x) • Radon gas from soil

Table 1. Sources of contamination in indoor environments.

As a result, lung diseases and allergies have become quite common, especially in young children and the elderly who often spend 90% of their time indoors, according to the Environmental Protection Agency.[2] The compound annual growth of asthma in North America between 2004 and 2008 is 4.9%, affecting more than 22 million people.[2] This chapter will describe a technology platform under development that seeks to balance indoor air quality requirements with energy consumption.

The three largest factors affecting energy consumption in commercial buildings, as illustrated in Fig. 2, are heating, ventilation and air conditioning systems (HVAC) (47%); lighting (23%); and office equipment (18%). The need for balance comes from the fact that preventing over-ventilation in times of low

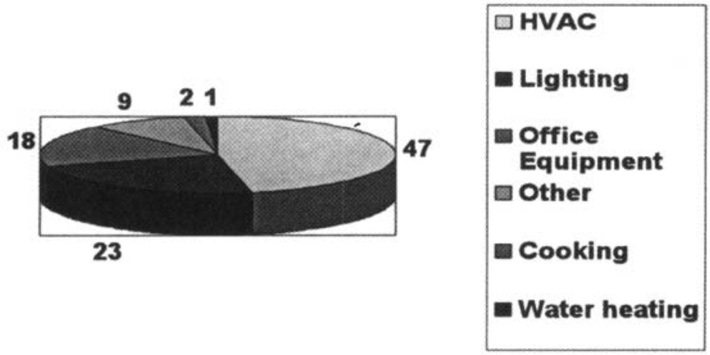


Figure 2. Sources of electric energy consumption.

occupancy also translates into high accumulation of emissions and deterioration of indoor air quality. Already, technologies such as motion sensing lighting, auto-dimming screens and residential lighting that dims and brightens in response to occupancy, are helping to reduce energy consumption.[3]

For example, carbon dioxide sensors are one of the most widely accepted and mature technologies on the market. When working with a demand controlled ventilation (DCV) system, CO_2 sensors can yield energy savings of 10–40% for a typical Californian city like Sacramento, where daily average temperatures can vary between 60 and 90 °F – see Table 2.[4] There is thus a clear incentive to consider adopting such DCV technologies.

It is useful to consider what type of sensors could provide a true indication of indoor air quality if DCV systems are adopted. Given performance, indoor air quality (IAQ) guidelines, and cost criteria, sensors for CO_2, relative humidity and formaldehyde are well suited for DCV control, with the formaldehyde sensor

	Savings
Daily average temperature (°C)	Sacramento, California (Bradshaw Road)
21.1	46%
26.7	28%
32.2	20%

Table 2. Energy savings as per use of demand controlled ventilation in a California study.

	Detection range	Response time	Resolution
Formaldehyde	25 – 250 μg/m^3	10 – 30 min	12 μg/m^3
All aldehydes	25 – 250 μg/m^3	10 – 30 min	12 μg/m^3
C-2 aldehydes	25 – 200 μg/m^3	10 – 30 min	12 μg/m^3
Toluene	100 – 5,000 μg/m^3	1 – 2 hrs	20 μg/m^3
Aromatic VOC	10 – 5,000 μg/m^3	1 – 2 hrs	10 μg/m^3
Total VOC	100 – 5,000 μg/m^3	1 – 2 hrs	20 μg/m^3
Benzene	5 – 100 μg/m^3	1 hr	5 μg/m^3
Radon	20 – 5,000 Bq/m^3	2 days	10 Bq/m^3

Table 3. Sensor specifications: detection range, response time and resolution.[5]

requiring the most development to decrease its unit price. If guideline values can be exceeded, then sensors for radon, total volatile organic compounds and particulate matter could also be integrated with a DCV control. Table 3 illustrates desirable sensor specifications for aldehydes and volatile organic compounds.[5] The conversion factor between μg/m^3 and ppb or parts per billion is: 1 ppb is equivalent to 1.2 μg/m^3.

For the researcher, the challenges lie in the current specifications of available sensors meeting higher concentrations of pollutants than what is desirable and likely to be legislated by health departments.

2. An integrated sensor network

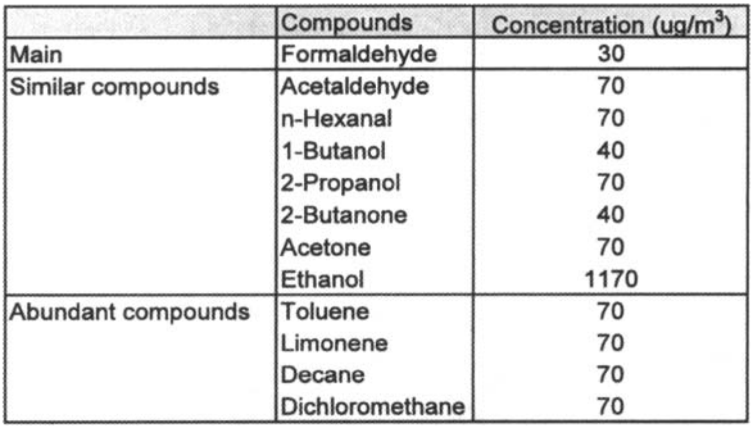
Like other groups worldwide, the intent of our NRC–ICT program is to develop an integrated indoor air quality sensor system with networking and closed loop HVAC actuation to monitor and control the indoor environment of commercial buildings, where improved measurements and responses would optimize both occupant health and comfort with minimum energy consumption.[6] This is to be achieved by:

- optimizing location and occupancy sensing;
- mining data to monitor equipment function and energy conservation; and
- developing a network of air quality sensors.

In addition to using sensors that are already on the market, research will be carried out to improve the sensitivity to aldehydes and VOCs, and the effectiveness of air sampling, data mining and actuation. The design goals are low cost, low power consumption, high sensitivity, high sampling rate, and continuous monitoring. Although sensing of toluene is as high a priority as that of formaldehyde, it is expected that a reassessment of current formaldehyde exposure guidelines by health authorities will come into effect sooner for formaldehyde than for toluene. The proposed new guidelines for formaldehyde will be short-term exposure (1 hour averaged) at 123 μg/m^3 (100 ppb) and long-term exposure (8 hours averaged) at 50 μg/m^3 (40ppb). Both the sensitivity at the parts per billion level and the selectivity pose substantial technical challenges to sensor development.

Health risks associated with exposure to radon have been reassessed based on additional scientific data and public consultations. Health authorities will recommend remedial measures when the average annual radon concentration exceeds 200 Bq/m^3 in normal occupancy area, with radon levels to be decreased to as low a value as practicable. Prolonged exposure to radon is the second highest cause of lung cancer in the United States.

Our experimental integrated sensor network is to be deployed in an indoor environment research facility with an experimental space of 12.2x7.3 meters providing a full-scale simulation of a commercial office setting. A dedicated HVAC system allows for control and monitoring of temperature, flow rate, humidity, pressure and percentage of outdoor air that is fed into the circulation.



	Compounds	Concentration (ug/m^3)
Main	Formaldehyde	30
Similar compounds	Acetaldehyde	70
	n-Hexanal	70
	1-Butanol	40
	2-Propanol	70
	2-Butanone	40
	Acetone	70
	Ethanol	1170
Abundant compounds	Toluene	70
	Limonene	70
	Decane	70
	Dichloromethane	70

Table 4. Characterization of formaldehyde sensors for sensitivity (detection range: 10–500 μg/m^3 and selectivity against a number of similar or abundant compounds.

In addition to the commercial off-the-shelf formaldehyde, temperature, relative humidity, radon and carbon dioxide sensors that will be integrated in the network, two chemical and two laser-based spectroscopic formaldehyde sensors will be developed, ensuring that the cross sensitivity to other aldehydes is smaller than the accuracy of the sensors. Table 4 illustrates the characterization of formaldehyde sensors for selectivity and sensitivity.

Factors that should be taken into consideration when selecting appropriate sensor elements have been summarized by Bishop and co-workers.[7] They include range, resolution, accuracy, precision, sensitivity, zero offset, linearity, zero drift, response time, signal to noise ratio, reproducibility to name a few.

3. Data mining

Although a lot of data exists on utility usage in private homes and commercial buildings, limited analysis has been done to dynamically control energy consumption in such environments. In the technology platform under development, the data management will include information from the air quality, HVAC, occupancy and energy systems as shown in Figure 3.

An illustration of preliminary data mining results in a controlled office environment where users' preferences for lighting are randomly chosen, is provided in Table 5. The data mining algorithm automatically determines the optimal desk/user assignments and dimming levels for the 9 lights, with a very low case discrepancy where

$$Discrepancy = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \; .$$

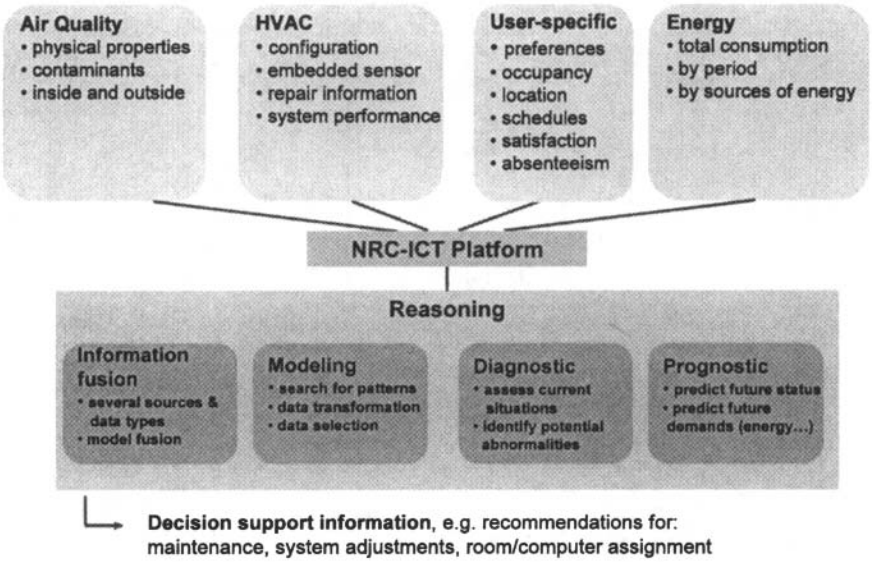


Figure 3. Data management for the integrated sensor network.

	Energy Consumption	Average Case Discrepancy (Lux)	Response Time (Seconds)
All-on model	100.0%	1454.44	0
All-off model	0.0%	1460.26	0
50%-on model	50.00%	490.13	0
Exhaustive search model	50.00%	0.0	20
Data mining model	**50.04%**	**4.93**	**2**

Table 5. Data mining algorithm for energy management taking into account users' preferences.

This preliminary analysis demonstrates that while realizing substantial energy conservation, the users can get a lighting environment that is very close to their personal preference with a short response time from the control system. The real challenge and ultimate value of a data mining function reside in the analysis of large quantity of data from numerous sources and the system optimization against criteria such as energy budget, life-cycle maintenance, healthy indoor environment, *etc.*

4. System overview

The NRC-ICT platform will incorporate a wireless communications network receiving a sensing measurement every minute. Eleven nodes are envisaged with an indoor transmission range of 15–30 meters and a capability of receiving data in a multi-floor environment. The communications protocol will co-exist with WLIS (zigbee) and Wi-Fi. A schematic of the platform is shown in Fig. 4.

Although first demonstrated in a lab environment, the indoor air quality sensor network system is intended to scale up for use in a multi-storey commercial building with its attendant challenges of retro-fitting and signal interference.

5. Conclusions

As more restrictive legislation comes into effect with regards to indoor air quality monitoring, the sensors for demand-controlled ventilation will succeed or fail based on three criteria: i) performance – detection range, resolution and stability; ii) cost – affordability; and iii) indoor air quality – typical indoor air quality levels exceed the guideline limits.[5]

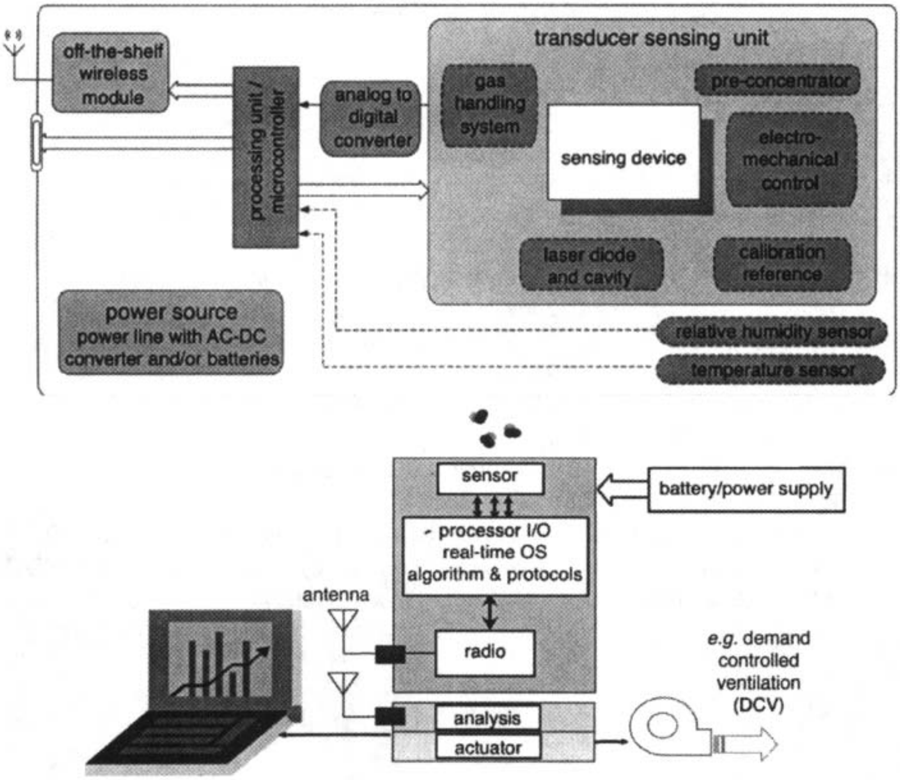


Figure 4. Sensor network system architecture and demonstration schematic.

As a greater awareness of the health hazards connected to physical environments permeates society, the need for fusion (network) of low-cost, robust IAQ sensors is the focus of current innovation in research laboratories both in the private and public sectors. Those innovations are enabled by multi-disciplinary approaches to optimizing sensor performance and integration into intelligent systems that control HVAC equipment. In addition to the well established requirement for physical and imaging sensors (*e.g.* infrastructure monitoring), there is a demand for chemical and biological sensors to be deployed in high traffic areas (*e.g.* airports, hospitals). Nanotechnologies are increasingly enabling the fabrication of ever smaller, low-power sensors that can be deployed over large areas (*e.g.* smart dust). Their use may be tempered by concerns regarding their toxicity and manufacturing hazards. Although the control of indoor environments according to occupancy and users' preferences may appear to be too invasive, the current reality is already one of ubiquitous cameras and monitoring.[8]

Acknowledgments

This work was supported by NRCC and its ICT Key Sector. Many thanks to the ICT Platform Development team and to Ruth Rayman, the ICT Key Sector coordinator, for providing materials for this chapter.

References

1. M. Conlin, "Is your office killing you?" *Business Week*, June 5th issue (2000).
2. Frost and Sullivan, "Strategic analysis of the North American indoor air quality markets," F381-19 (2005).
3. J. Heimbuch, "Panasonic's auto-dimming lights for energy efficiency," *Tree Hugger*, January 30 issue (2009).
4. I. Krepchin, C. Scruton, and J. Braun, 'Commercial buildings breathe right with DCV," *Eng. Syst.* **23**, 38 (2006).
5. D. Won and W. Yang, "The state-of-the-art in sensor technology for demand-controlled ventilation," Report IRC-RR-243 (2005).
6. NRC–ICT Key Sector Office, "Indoor air quality sensor network – system design document," (2009).
7. R. H. Bishop, *The Mechatronics Handbook*, Boca Raton: CRC Press, 2002.
8. Ontario Center for Environmental Technology Advancement (OCETA), "Market assessment for environmental sensors," (2007).

Quantum-Dot Infrared Photodetectors: In Search of the Right Design for Room-Temperature Operation

A. Sergeev, V. Mitin, L. H. Chien, and N. Vagidov
Dept. of Electrical Engineering, SUNY–Buffalo
Buffalo, NY 14260, U.S.A.

1. Introduction

Development of sensitive low-cost far-infrared photodetectors is critical for various applications in public safety, industry, healthcare, and defense technologies.[1,2,3] Sensitive detectors operating at the room temperature would significantly increase both commercial and defense markets, where the quantum-well infrared photodetectors (QWIPs) currently prevail over other sensors. The advantages of QWIPs include low cost, high uniformity and reproducibility due to well-established GaAs growth and processing technologies. Large area QWIP arrays are employed in imaging devices operating at liquid nitrogen temperatures and below.[4] Modern QWIPs demonstrate the detectivity of $\sim 10^{10}$ cm$\cdot$Hz$^{1/2}$/W for λ = 10 μm at 77 K. However, the detectivity drops to 10^8 cm$\cdot$Hz$^{1/2}$/W at room temperature. It is well understood that the high-temperature limitations of QWIPs and many other photodetectors are caused by a tremendous decrease of the photocarrier lifetime, which strongly reduces the responsivity and sensitivity.[5]

Quantum-dot infrared photodetectors (QDIPs) were proposed by Ryzhii[6] in 1996. Since then, QDIPs have been considered the most promising candidate for achieving room-temperature operation. Initially, these hopes were pinned on the long photoexcited electron lifetime associated with the so-called "phonon bottleneck" (for a review see Ref. 1). In this simple theoretical picture, one ignores the finite width of electron energy levels and assumes that only specific phonons are excited due to electron intra-dot transitions.[7] The small phase volume of these phonons should then substantially suppress the electron relaxation. However, experimentally measured phonon-mediated carrier relaxation turned out to be much faster than expected in the phonon bottleneck model.[8,9] After numerous experiments with various QD structures, the QDIP community concluded that in the IR range "there is no true phonon bottleneck".[1]

During the last few years, QDIPs have attracted significant scientific interest due to various intriguing possibilities for engineering QD structures with specific properties. After intensive theoretical and experimental research, several groups[10,11] optimized the traditional QDIP design and demonstrated room-temperature operation with detectivities $D^* \sim 10^7$–10^8 cm$\cdot$Hz$^{1/2}$/W, which are close to detectivities of QWIPs. A leader of one of these groups summarized the further perspectives for QDIPs as follows:[1] "there is still room for improvement of QDIP

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

detectivities, though perhaps not as much as one would like, at least in this particular current geometry".

In this chapter, we consider some novel ideas in QDIP design and identify the physical mechanisms that can effectively control photoelectron kinetics in QDIPs. We focus on the effect of potential barriers around individual dots, dot planes, and dot clusters. Such barriers can effectively separate the localized intra-dot electron states and conducting states in the matrix. By engineering these potential barriers, we can substantially suppress the photocarrier capture processes and in this way improve the QDIP performance.

2. Ultimate QDIP performance

The ultimate noise characteristics of a detector for a specific spectral range are determined by fluctuations of the background radiation. In a regime of the background-limited performance (BLIP), the intrinsic noise is equal to the background noise. If the photon flux absorbed by a detector of the area S is $\eta_0\Phi_{BG}S$, the average fluctuation is given by $\eta_0(\Phi_{BG}S)^{1/2}$, where Φ_{BG} is the background flux density and η_0 is the absorption coefficient. Introducing the intrinsic quantum efficiency η_i as the probability of carrier generation by an absorbed photon, we obtain $\eta_i\eta_0(\Phi_{BG}S)^{1/2}$ as the fluctuation in the photocarrier generation rate. The intrinsic generation-recombination noise is determined by the kinetics of thermal carriers. Near thermodynamic equilibrium, the generation rate is $n_{th}V/\tau$ and its average fluctuation is given by $(n_{th}V/\tau)^{1/2}$, where V is the detector volume, n_{th} is the thermally-generated carrier concentration and τ is the carrier lifetime. In the BLIP regime, thermal generation noise should be as low as the photogeneration noise due to background fluctuations, *i.e.* $(n_{th}V/\tau)^{1/2}$ = $\eta_i\eta_0(\Phi_{BG}S)^{1/2}$. Thus, the detector parameters in the BLIP regime are related as:[12]

$$\Phi_{BLIP} = \frac{n_{th}d}{(\eta_0\eta_i)^2\tau} = G_{th} \quad , \tag{1}$$

where $d = V/S$ is the detector thickness. Note, that in most theoretical papers[13,14] on QDIP detectors, the analogous equation was derived with the quantum efficiency $\eta = \eta_0\eta_i$ in the denominator of Eq. (1), instead of $\eta^2 = (\eta_0\eta_i)^2$. In our strong opinion, such consideration of noise processes is incorrect, because absorption does not change the relative fluctuation in the photon flux.

The noise processes in a detector are usually described in terms of the noise equivalent power (NEP), which is defined as an equivalent fluctuation of the background radiation power. Thus, the NEP and detectivity determined by the generation-recombination noise are given by:

$$\mathrm{NEP}_{GR} = \frac{h\nu}{\eta}\sqrt{\frac{2n_{th}V}{\tau}} = h\nu\sqrt{2SG_{th}} \ , \quad D^* \equiv \frac{\sqrt{S}}{\mathrm{NEP}_{GR}} = \frac{\eta}{h\nu}\sqrt{\frac{\tau}{2n_{th}d}} \ . \tag{2}$$

In QDIPs the thermally generated carrier density may be expressed in terms of the equilibrium electron concentration n_0 of localized states in the dots as follows:

$$n_{th} = n_0 g_S \exp[(E_C - E_L)/kT] \,, \tag{3}$$

where E_C and E_L are the energy of conducting and localized electron states correspondingly, and g_S is the factor of electron statistics. For an optimally designed device, the photon energy $h\nu$ should just slightly exceed $(E_C - E_L)$.

Modern QDIPs intended for room-temperature operation have $n_0 d \sim 10^{15}/\mathrm{cm}^2$, $g_S \sim 1$, and η up to 0.3–0.4. The carrier lifetime τ in traditional QDIP structures is of the order of 1–10 ps at room temperature. Substituting these parameters in Eqs. (2) and (3), yields a detectivity of $\sim 10^8$ cm·Hz$^{1/2}$/W for $\lambda = 10$ μm, in agreement with recent experimental results. This means that further QDIP improvement will require new ideas and significant changes in the detector design.

3. Potential barriers in QDIP structures

The photocarrier capture processes can be suppressed by employing specially-engineered structures with potential barriers around dots. Such barriers are created by electrons bound in quantum dots and ionized impurities in the depletion areas. These barriers separate conducting electron states in the matrix from the localized intra-dot states and, in this way, prevent photoelectrons from being captured into dots. Modern technologies provide many opportunities to fabricate QDIPs with local and collective potential barriers. We take local barriers to mean barriers around a single dot; whereas collective barriers to surround groups of quantum dots (*e.g.* the dot rows in Fig. 1) and separate them from high-mobility conducting channels. With the combination of local and collective barriers, it is possible to create very specific structures with strongly suppressed processes of electron capture and enhanced electron coupling to infrared (IR) radiation. In our opinion, the optimal structures with collective barriers will combine advantages of QWIPs (such as strong coupling to radiation due to large number of electrons in localized states) with manageable kinetics of QD structures. In general, photocarriers in QDIPs can overcome the barriers via tunneling and thermo-excitation processes. In structures with local barriers, both processes can be important depending on the dot and matrix parameters.[15] In structures with collective barriers, the tunneling mechanism is strongly suppressed due to large barrier width.

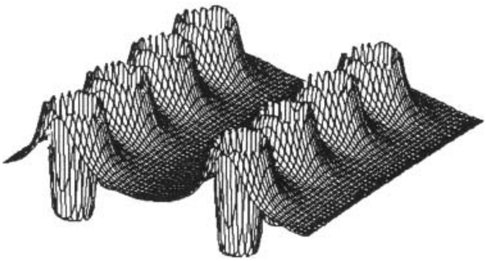


Figure 1. Band structure of QD rows.

In our previous publications,[12,15,16] we considered structures with local barriers. Our main conclusions may be summarized as follows. The photocarrier lifetime critically depends on three factors: the ratio between the characteristic dot radius a and the electron mean free path l; the height of the potential barrier V_M; and the heating of photocarriers by the electric field E. In the low electric field limit, if $l > a$, the photoelectron lifetime is given by: [12,16]

$$\frac{1}{\tau} = \pi N_D a^3 \frac{1}{\tau_\varepsilon} \exp(-eV_M/kT) \quad , \tag{4}$$

where N_D is the dot concentration and $1/\tau_\varepsilon$ is the inelastic electron-phonon relaxation rate. The exponential factor describes the effect of potential barriers on capture processes.

In strong electric fields, on the other hand, photocarrier kinetics and transport are well described by the electron heating model. Our simulations show that the barriers strongly suppress capture processes until the average energy of carriers, *i.e.* the electron temperature, becomes comparable with the barrier height.[12,16]

To effectively suppress photoelectron capture, the barrier height V_M should be at least 2–3kT. Such a local potential at room temperature means a dot should contain ten or more electrons.[15] This requirement creates a lot of practical problems, such as high doping levels, specific large dot sizes, *etc.* Below, we shall focus on creating effective collective barriers in lateral and vertical structures.

4. QDIPs with lateral transport

A schematic QDIP with lateral photocarrier transport is shown in Fig. 2(a). A typical photodetector is based on the modulation-doped AlGaAs/GaAs with self-assembled InAs QDs. The QD layers are embedded in GaAs quantum wells of thickness of $2d$, while the AlGaAs barriers are doped to supply the dots with electrons or holes.

A typical detector might contain ~10 QD layers to match the wavelength of interest. Ohmic contacts to the lateral QD structures are fabricated by depositing and annealing AuNiGe alloy. Because of the complex contact technology, the inter-electrode distance is relatively large: $L \sim 100$ µm. The corresponding band diagram of the structure is shown in Fig. 2(b). Potential barriers are created by the

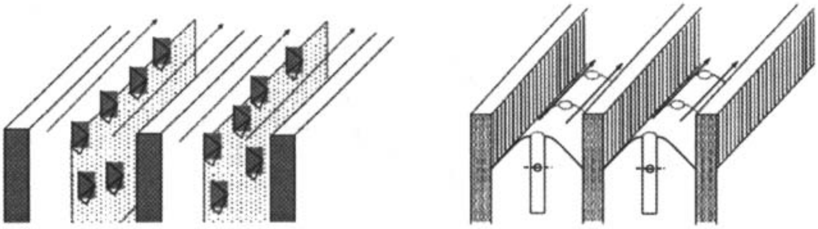


Figure 2. The QDIP with the lateral transport (a) and the corresponding band structure (b).

charged QD planes and charged planes of dopants. Photoexcited carriers move along areas with high mobility near the modulation-doping heterointerfaces. Thus, potential barriers between high-mobility conducting states and localized states in QDs are proportional to the distance *d* between the QD plane and AlGaAs barrier.

The photoelectron kinetics in such lateral structures was experimentally investigated at T = 77 K,[17] with the result that the photocarrier lifetime and the detector responsivity depended exponentially on *d*. The lifetime was measured to be as long as 3 ms. The corresponding photoconductive gain g_{PH} was evaluated as ~10^6. In this way, the critical importance of the collective potential barriers in lateral structures was fully confirmed by measurements at T = 77 K. At higher temperatures (~200 K), the QDIP performance deteriorated markedly because of the parasitic effects in Ohmic contacts.

We now turn to the theoretical performance limits of the lateral QDIPs at the room temperature by modeling photocarrier capture. Our Monte-Carlo program includes electron scattering on acoustic, polar optical, and intervalley phonons. We use the Γ–L–X model for simulation kinetics and transport in GaAs/AlGaAs structures with InAs dots.

Our results show that in small electric fields, the capture time exponentially increases with increasing of the height of the potential barrier V_M. In the lateral structures, the potential is created by charged planes and it may be presented as $V_M = e^2 N_D n d/(2\varepsilon_r \varepsilon_0)$ (see the inset in Fig. 3), where N_D is the dot concentration in QD planes, and *n* is the dot occupation, *i.e.* the average number of carriers per dot. Thus, the capture time increases exponentially with dot concentration, occupation, and the well width *d*. In Fig. 3(a) we present the dependence of the capture time on the occupation *n* for typical parameters: $N_D = 4\text{x}10^{10}$ cm^{-2} and d = 80 nm. As in the case of QD structures with local barriers, the effect of the electric field can be understood in terms of the electron heating. The electric field increases the average energy (temperature) of carriers, so the barriers are less effective in preventing the capture of hot electrons. Therefore, the capture time decreases as the electric field increases. The detector responsivity is proportional to the photoconductive gain g_{PH}, which is defined as a ratio of the photoelectron lifetime

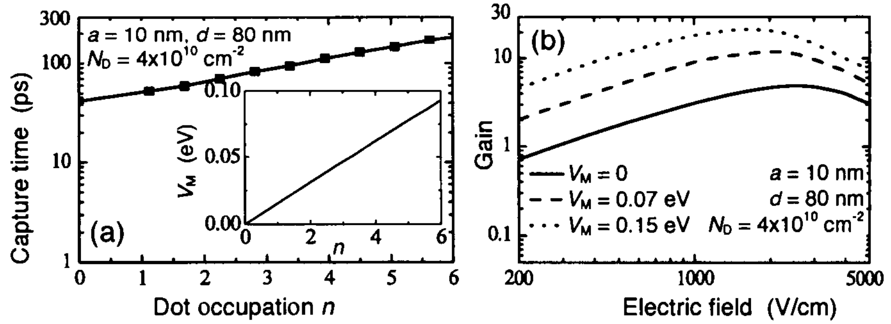


Figure 3. (a) Capture time *vs.* dot carrier occupation *n*, barrier height V_M is shown in the inset; (b) photoconductive gain *vs.* electric field.

to the transit time ~ $L/\mu E$. Figure 3(b) plots the dependence of the photoconductive gain on the electric field. The initial increase of g_{PH} with E is due to the decrease of the transit time. The decrease of g_{PH} in strong fields is due to a decrease in the capture time. Our modeling demonstrates that collective barriers increase the photoelectron lifetime and photoconductive gain in the room-temperature lateral QDIPs. While for the structures used in Ref. 17 this increase is not predicted to be sufficient for room-temperature applications, the effect can be significantly enhanced in QD structures with higher dot concentrations and larger values of the well width d.

5. QDIPs with vertically correlated dot clusters

The advantages of QWIPs with the vertical transport over the lateral QWIPs include a shorter intercontact distance and well-established processing technologies that promise high-quality reproducible Ohmic contacts. The short detector base decreases the generation-recombination noise and increases the photoconductive gain. High-quality Ohmic contacts are critical for both the detector responsivity and noise characteristics. To combine the advantages of the vertical QDIPs with barrier-limited capture, we propose a detector based on structures with vertically correlated dot clusters (VCDC).

A schematic of the VCDC structure is presented in Fig. 4. Positions of quantum dot clusters are correlated in the direction of the electric current, *i.e.* in the vertical direction. Thus, the correlated clusters of QDs form pancake-like structures. The collective barriers around QD clusters are created by charged dots. These barriers separate the conducting channels in the matrix from the localized states in quantum dots. Having been excited by radiation, photocarriers move in the areas between dot clusters through high-mobility conducting channels, as shown in Fig. 4. If the cluster radius b exceeds the distance between dot planes c, the barrier potential around clusters has a logarithmic form $V_M = e^2Nn/(2\pi\varepsilon_r\varepsilon_0 c) \ln(w/b)$, where N is the number of dots in the cluster, n is the average number of carriers per dot, and $2w$ is the distance between the centers of two clusters.

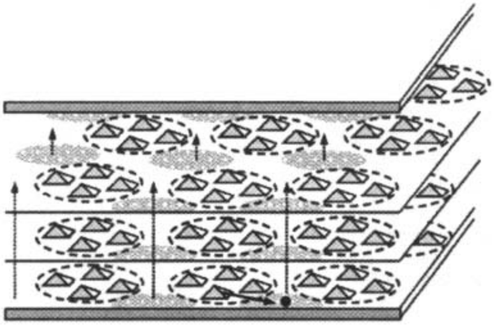


Figure 4. Structure with vertically correlated dot clusters (VCDCs).

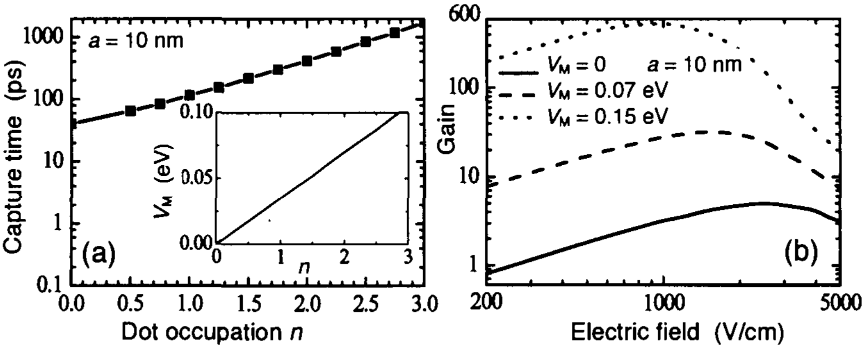


Figure 5. (a) Capture time and the barrier height (inset) *vs.* dot occupation *n*; (b) photoconductive gain *vs.* electric field.

Results of the Monte-Carlo modeling of photocarrier kinetics in VCDC structures show that the capture time exponentially increases with increasing of the number of dots in a cluster N and the occupation n. At the same time, the capture processes are weakly sensitive to geometrical parameters w and b. Figure 5(a) illustrates the exponential dependence of the capture time on the occupation of dots for $N = 9$, $b = 75$ nm, and $w = 150$ nm. We also investigate the dependence of the capture time on the electric field E, which is presented in Fig. 5(b). Again, this dependence is nonmonotonic due to electron heating effects.

Thus, the modeling shows that compared with ordinary quantum-dot structures, where the photoelectron lifetime at room temperatures is of the order of 1–10 ps, the VCDC structure with collective potential barriers would allow for increasing the photocarrier lifetime by up to three orders of magnitude.

6. Conclusions

Long photocarrier lifetime is a key issue for improving of room-temperature infrared photodetectors. Detectors based on quantum dot structures permit the engineering of specific properties and hence hold promise of overcoming the limitations of quantum well detectors. In this chapter we reviewed photocarrier kinetics in traditional QDIPs and presented the modeling of the capture processes in the lateral structures and vertical structures with vertically correlated dot clusters. The results show that QDIPs can substantially outperform QWIPs due to manageable photocarrier kinetics, which can be controlled by potential barriers created by dot clusters. Modern technologies allow the fabrication of various structures with the correlated QD clusters. Detailed modeling of such structures will pave the way for optimal design of the room-temperature QDIPs.

Acknowledgments

This work was supported by an AFOSR contract.

References

1. M. Razeghi, H. Lim, S. Tsao, J. Szafraniec, W. Zhang, K. Mi, and B. Movaghar, "Transport and photodetection in self-assembled semiconductor quantum dots," *Nanotechnology* **16**, 219 (2005).
2. J. C. Campbell and A. Madhukar, "Quantum dot infrared detectors," *IEEE J. Quantum Electronics* **95**, 1815 (2007).
3. A. V. Barve, S. J. Lee, S. K. Noh, and S. Krishna, "Review of current progress in quantum dot infrared photodetectors," *Laser & Photonics Reviews*, June 17, 2009.
4. B. F. Levine, "Quantum-well infrared photodetectors," *J. Appl. Phys.* **74**, R1 (1993).
5. E. A. Zibik, T. Grange, B. A. Carpenter, *et al.*, "Long lifetimes of quantum-dot intersublevel transitions in the terahertz range," *Nature Mater.* **8**, 803 (2009).
6. V. Ryzhii, "The theory of quantum-dot infrared phototransistors," *Semicond. Sci. Technol.* **11**, 759 (1996).
7. U. Bockelmann and G. Bastard, "Phonon scattering and energy relaxation in two-, one-, and zero-dimensional electron gases," *Phys. Rev. B* **42**, 8947 (1990).
8. Y. Toda, O. Moriwaki, M. Nishioka, and Y. Arakawa, "Efficient carrier relaxation mechanism in InGaAs/GaAs self-assembled quantum dots based on the existence of continuum states," *Phys. Rev. Lett.* **82**, 4114 (1999).
9. R. Ferreira and G. Bastard, "Phonon-assisted capture and intradot Auger relaxation in quantum dots," *Appl. Phys. Lett.* **74**, 2818 (1999).
10. P. Bhattacharya, X. H. Su, S. Chakrabarti, G. Ariyawansa, and A. G. U. Perera, "Characteristics of a tunneling quantum-dot infrared photodetector operating at room temperature," *Appl. Phys. Lett.* **86**, 191106 (2005).
11. S. Tsao, H. Lim, W. Zhang, and M. Razeghi, "High operating temperature 320×256 middle-wavelength infrared focal plane array imaging based on an InAs/InGaAs/InAlAs/InP quantum dot infrared photodetector," *Appl. Phys. Lett.* **90**, 201109 (2007).
12. L.-H. Chien, A. Sergeev, N. Vagidov, and V. Mitin, "Hot-electron transport in quantum-dot photodetectors," *Intern. J. High Speed Electronics Syst.* **18**, 1013 (2008).
13. J. Phillips, "Evaluation of the fundamental properties of quantum dot infrared detectors," *J. Appl. Phys.* **91**, 4590 (2002).
14. P. Martyniuk and A. Rogalski, "Quantum-dot infrared photodetectors: Status and outlook," *Prog. Quantum Electronics* **32**, 89 (2008).

15. A. Sergeev, V. Mitin, and M. Stroscio, "Quantum-dot photodetector operating at room temperatures: Diffusion-limited capture," *Physica B* **316-317**, 369 (2002).
16. V. Mitin, A. Sergeev, L.-H. Chien, and N. Vagidov, "High performance of IR detectors due to controllable kinetics in quantum-dot structures," *Proc. SPIE* **7095**, 70950K (2008).
17. K. Hirakawa, S.-W. Lee, Ph. Lelong, *et al.*, "High-sensitivity modulation-doped quantum dot infrared photodetectors," *Microelectronic Eng.* **63**, 185 (2002).

This page intentionally left blank

Treating the Case of Incurable Hysteresis in VO_2

M. Gurvitch
Dept. of Physics and Astronomy, SUNY–Stony Brook, NY 11794 and
NY State Center for Advanced Sensor Technology, Stony Brook, NY 11794, U.S.A.

S. Luryi
Dept. of Electrical and Computer Engineering, SUNY–Stony Brook and
NY State Center for Advanced Sensor Technology, Stony Brook, NY 11794, U.S.A.

A. Polyakov and A. Shabalov
NY State Center for Advanced Sensor Technology, Stony Brook, NY 11794, U.S.A.

1. Introduction

Phase transitions present an opportunity for a useful application utilizing a natural proximity of two phases with different physical properties and an ability to trigger the transition by varying a transition-controlling parameter, such as temperature, or magnetic field, *etc*. There exist applications based on magnetic transitions, optical transitions, metal-insulator transitions, and superconducting transitions. Some of these phase transitions, especially those classified as *first-order* transitions (*i.e.* accompanied by the release or absorption of latent heat), are hysteretic in nature, so that the two phases transform into each other along different routes depending on the direction of change of a transition-controlling parameter. Such hysteresis can be essential for an application. A well-known example (in this case, of a hysteretic *second-order* transition) is a ferromagnetic material that can be magnetized by an external magnetic field and, because of the hysteresis, will remain magnetized when the external field is reduced to zero. This is used for making permanent magnets and in magnetic storage of information. The stored, "memorized" magnetization can be erased, which requires specific steps to be taken, such as going around the magnetic hysteresis loop with diminishing external fields. In other cases, however, the utilization of a phase transition is hindered by its hysteresis. The "memory" of the previous history that is beneficial in magnetic materials may become detrimental in other applications, where the irreversible (or, more precisely, not easily reversible) change in a given material's property, *e.g.* resistivity, may be undesirable.

In this chapter we consider VO_2, a material undergoing a first-order phase transition, and target an application – infrared (IR) visualization utilizing resistive bolometers made of VO_2 – in which hysteresis causes problems. We offer a way of operation that circumvents these problems. The applied aspects of IR visualization vis-à-vis our proposed approach are discussed in our recent

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky
Copyright © 2010 John Wiley & Sons, Inc.

publication,[1] and we will refer to it in a number of instances for further details and references. Here we want to explain the essence of the method and touch upon its physics fundamentals.

2. Hysteretic semiconductor-metal phase transition in VO_2 – major and minor loops

The resistivity of VO_2 changes by 3–5 orders of magnitude in a spectacular semiconductor-to-metal phase transition around 68 °C.[2] In addition to resistivity, other properties change in this first-order phase transition, including crystalline structure and optical constants. In single crystals, the transition is very sharp, with hysteresis width of 0.5–2 °C,[3] whereas in polycrystalline films it is broader, with hysteresis width (as measured in the middle of a transition) of 10–20 °C. The resistive transition measured in one of our VO_2 films can be seen in Figs. 1-4 (look at the outside *major* loop; ignore for now the inner or *minor* loops or their parts which are shown on these figures). The films used in these measurements were prepared on Si/SiO_2 substrates by pulsed laser deposition; details of sample preparation and measurements can be found in Ref. 1. On the log(R) *vs.* T plot starting from room temperature, the resistivity first follows the semiconducting S-phase slope with increasing T and then falls sharply along the right (heating) branch of the transition reaching metallic M-phase. When the temperature is decreased, the resistivity follows the left (cooling) branch, forming a closed loop between two transition-end points T_S and T_M. As can be seen in Figs. 1-4, the resistivity below the transition, in the S-phase, is about 3 orders of magnitude

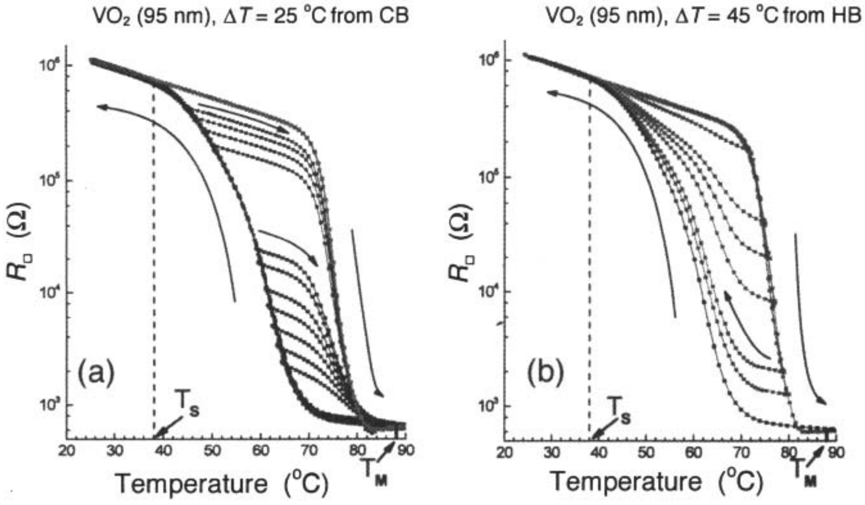


Figure 1. Major hysteresis loop of VO_2 film with a number of minor loops initiated from various temperatures T_0 on (a) the cooling branch (CB), $T_0 \rightarrow T_0+\Delta T \rightarrow T_0$ with $\Delta T = 25$ °C, and (b) the heating branch (HB), $T_0 \rightarrow T_0-\Delta T \rightarrow T_0$ with $\Delta T = 45$ °C.

higher than in the M-phase above the transition. The hysteresis width in the middle of this log(R) *vs.* T plot is about 15 °C, while the total width of a hysteretic region is $T_M - T_S \sim 50$ °C.

For a variety of reasons, despite the many proposals to use this transition in applications, almost none materialized. One commercial and military application in which the use of VO_2 was initially envisioned is IR visualization (night vision) based on resistive microbolometers. However, technology eventually settled on a "poor cousin" of VO_2, a non-transitioning mixed VO_X oxide, mainly to avoid hysteresis, which greatly complicates bolometer operation, as was mentioned in the introduction and will be discussed in more detail below.

The major hysteretic loop encompasses all the points on the (T, R) plane that can serve as starting points of minor hysteretic loops. In particular, minor loops can be initiated at any point along the outer lines of the major loop. In Fig. 1, we show several minor loops which were traced by reversing the temperature change at various points along the major loop: in Fig. 1(a), minor loops are starting from points along the cooling branch (CB); in Fig. 1(b), from points along the heating branch (HB). The loops shown in Fig. 1 cover wide temperature ranges of 25 °C in Fig. 1(a) and 45 °C in Fig. 1(b). A number of minor loops traced over a narrower temperature range of 10 °C are shown in Fig. 2.

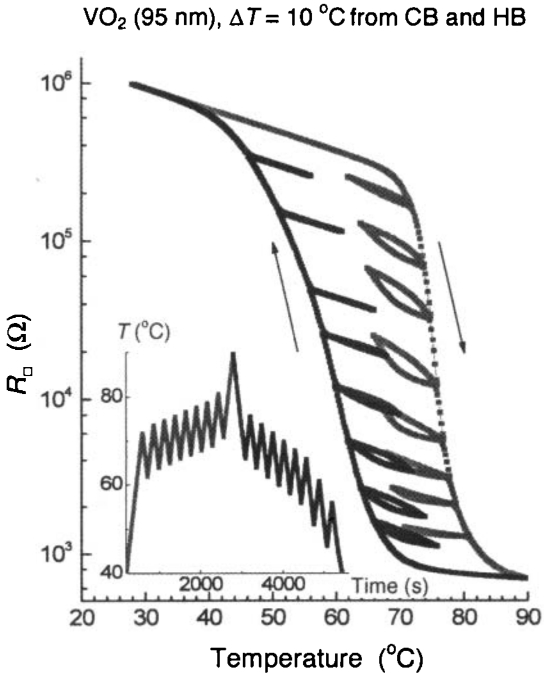


Figure 2. Minor loops with $\Delta T = 10$ °C, the inset shows the way in which temperature was changed in this measurement.

3. How hysteresis causes problems in bolometric readout – forward and backward excursions

A bolometer reacts to a temperature change ΔT by changing its resistance by ΔR. The larger the ΔR for a given ΔT, the greater is the sensitivity. At the same time, very large resistance R is detrimental because of bolometric sensor Joule heating during readout, the difficulty of matching to the electronic readout circuit, and higher noise, both Johnson and $1/f$.[1] Thus, a commonly-used measure of the bolometric material's sensitivity is the logarithmic derivative $(1/R)(\Delta R/\Delta T)$ known as the temperature coefficient of resistance, or TCR. In other words, it is not the high ΔR, but high $\Delta R/R$ that is desired.

The steep semiconductor to metal transition seems to promise great sensitivity (high TCR), comparable to the sensitivity that can be achieved in a transition-edge superconducting sensor, which provided the original incentive to employ such a transition in VO_2.[4] Why then was it not used? The main reason is the hysteresis.

Let us look in more detail on how hysteresis causes problems. To be specific, let us consider a bolometer positioned at a working point (T_0, R_0) on a steep heating branch (HB), kept there using a temperature controller as T_0 is above room temperature. Then, an influx of energy (this could be an IR signal) momentarily heats it up, increasing its temperature by ΔT and decreasing its resistance by a large ΔR, as shown in Fig. 3

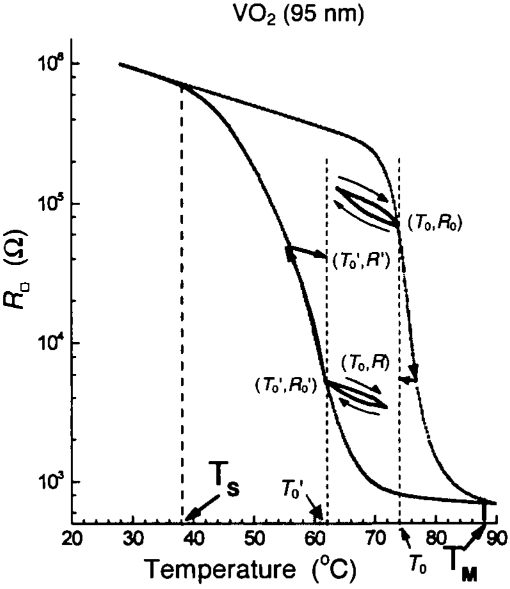


Figure 3. Forward excursions originating from point (T_0, R_0) on an HB and from point (T_0', R_0') on a CB, tracing open curves ending at (T_0, R) on an HB and at (T_0', R') on a CB. The backward excursions from (T_0, R_0) and (T_0', R_0') return to the points of their origin following closed minor loops. Note that this figure is not a schematic drawing; it shows measured loops and parts of loops.

Once the heat input has been removed (which happens in resistive bolometers used in IR visualization at least 30 times per second), the bolometer returns back to the temperature T_0; however, because of the hysteresis, the (T, R) path on the way back differs from the path in the forward direction: the resistance will not come back to R_0, but instead will move along the part of the *minor* loop to a point (T_0, R) as indicated in Fig. 3. If a bolometer now experiences a subsequent second heat pulse, the result will be very different. This cannot be tolerated in a bolometric sensor subject to a train of pulses. In order to avoid this problem the bolometer should be reset after each pulse, which can only be done by going all the way out of the hysteresis loop, reaching temperatures either below T_S or above T_M, as can be seen in Fig. 1. While such resetting can be done, it requires large temperature *excursions* (we will call a round-trip temperature change $T_0 \rightarrow T_0 \pm \Delta T \rightarrow T_0$ an *excursion*) followed each time by temperature stabilization at the working point T_0. With bolometers receiving signals at 30–60 Hz (video rates), this is not practical. The problem with the different return path we just outlined has been described in the literature,[5] and the readout problems caused by hysteresis were acknowledged[6] as the major contributing factor in killing the original proposal[4] of using high TCR found in a VO_2 transition.

A word about terminology: as was said, we call a round trip $T_0 \rightarrow T_0 \pm \Delta T \rightarrow T_0$ an *excursion*. We will take the *excursion* length ΔT to be a positive quantity, and show the plus or minus sign in front of it explicitly. If temperature of the initial part of an excursion changes in the same direction as the major loop progression, we call it a *forward excursion*; if it starts in the opposite direction, it is a *backward excursion*. Thus, in Fig. 3 on the HB, the excursion from (T_0, R_0) to (T_0, R) is a forward one; the other one, forming a closed loop, is a backward one. We could discuss similar processes originating from point (T_0', R_0') on a CB (they are also shown in Fig. 3), keeping in mind that on a CB decreasing temperature corresponds to the forward direction, and increasing temperature to the backward direction.

Because of the hysteresis, forward and backward excursions produce dramatically different results. While a forward excursion $T_0 \rightarrow T_0+\Delta T \rightarrow T_0$ on a HB traces an *open* curve from (T_0, R_0) to (T_0, R), a backward excursion $T_0 \rightarrow T_0-\Delta T \rightarrow T_0$ produces a *closed* minor loop, which has a much smaller slope near (T_0, R_0). The slope increases with the excursion ΔT, thereby introducing a nonlinearity in $\log(R)$ *vs.* T – the TCR now depends on the value of ΔT. Similar processes take place on a CB, see Fig. 3.

To summarize, hysteresis causes the following two distinct types of problems:

- forward excursions produce open (T, R) curves, a "memory effect" that is unacceptable in a bolometer taking multiple repeated measurements;
- backward excursions produce closed loops, so that bolometer re-setting is not required. However, the double-valued nature of such a minor loop makes readout ambiguous, and variable TCR found in minor loops is unacceptable, or at least detrimental, producing nonlinearities in the bolometric response, with larger ΔT corresponding to larger TCR.

4. Nonhysteretic branches in resistivity

Now let us examine Fig. 2 in more detail. We see that most of the minor loops in Fig. 2 are hysteretic, but we also notice that some of them are rather flat and nearly single-valued, particularly those near the major loop ends T_S and T_M. In studying minor loops with progressively smaller excursions ΔT we discovered that for sufficiently small ΔT minor loops flatten out, degenerating into what we call *nonhysteretic branches* (NHBs). Although some minor loops become flat with ΔT = 10 °C (Fig. 2), *all of them* become flat at or below ΔT_{NHB} = 4–5 °C, as can be seen in Fig. 4.

An NHB can be initiated from any point on the major loop, either on a HB or on a CB; NHBs are linear in log(R) *vs.* T with negative slopes (semiconducting TCRs). Two representative NHBs measured in round-trip excursions are plotted in Fig. 5.

The NHBs are single valued to the precision of our measurements, except for the temperature intervals near the attachment point T_0 and the turning point $T_0 \pm \Delta T$, where they sometimes appear double-valued, exhibiting a small loop and a fork, as seen in Fig. 5(a). These features depend on the rate of temperature sweep; they are probably instrumental effects resulting from a lag between film surface and thermometer temperatures.[1]

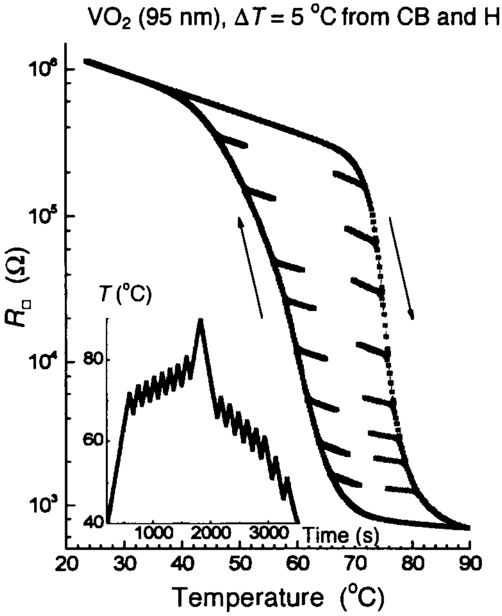


Figure 4. Minor loops with ΔT = 5 °C; inset shows the temperature variation during this measurement.

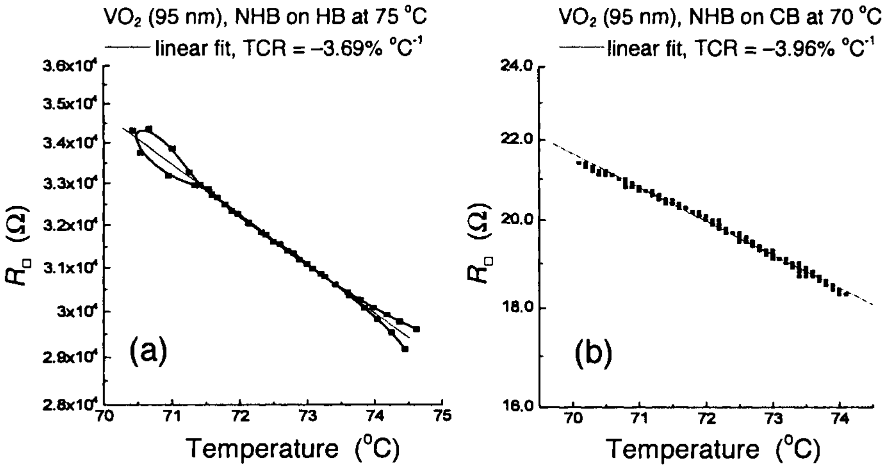


Figure 5. (a) NHB attached to the HB, with $R_\square \approx 31$ kΩ; TCR = −3.69 % °C^{-1}; (b) NHB attached to the CB, with $R_\square \approx 20$ Ω; despite this low R, TCR = −3.96 % °C^{-1}.

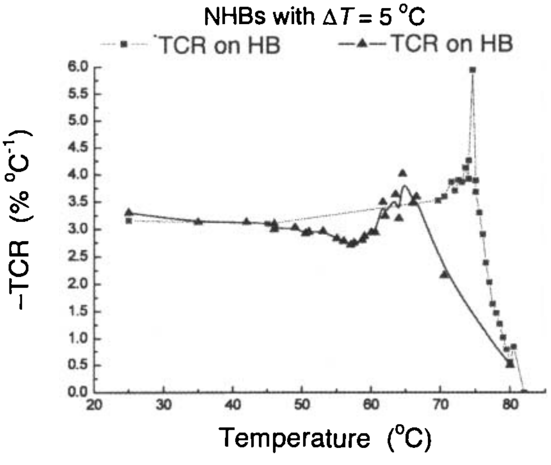


Figure 6. Absolute values of TCR *vs.* T_0 for various NHBs around the major loop.

The TCRs of various NHBs from Fig. 4 are plotted *vs.* T_0 in Fig. 6 for both HB and CB. The S-phase TCR at 25 °C is 3.3 % °C^{-1}, which is considerably higher than a typical TCR ≈ 2 % °C^{-1} at 25 °C found in the literature on VO_X.[1,6] In an intrinsic semiconductor, the bandgap E_G is given by[1] $2(|\mathrm{TCR}|)k_BT^2$. Substituting $|\mathrm{TCR}| = 0.033$ and $T = 25$ °C = 298 K, we obtain $E_G = 0.51$ eV, in fair agreement with the bandgap value in VO_2 single crystals.[3]

As we see in Fig. 6, NHBs logarithmic slopes peak at 65 °C on the CB and 75 °C on the HB, with maximum values of 4% °C^{-1} and 6% °C^{-1}, decreasing sharply above the peaks. We note that the highest TCRs on the sides of the major loop are found at these same temperatures.

5. Theoretical model

We will present a theoretical model that provides a qualitative explanation of the observed behavior and has predictive power that will be put to a test in the next Section.

The hysteretic region in VO_2 is a mixed state consisting of both the S-phase and M-phase regions, or domains. Each such region located in a film around a point with spatial coordinates (x, y) transitions into the other phase at its own temperature T_C, with the variation in T_C arising from nonuniformities of composition, variations in the local strain, *etc.* In a macroscopic sample $T_C(x, y)$ is quasi-continuously distributed. We assume that the local transition within a domain is sharp; further, we assume that a uniform isolated domain would transition without a hysteresis. In this we differ from much of the VO_2 hysteresis literature[7] in which it is usually assumed that each domain, in addition to a local T_C, has its own coercive temperature and a rectangular hysteresis loop. We note that single crystals have very small hysteresis;[3] it seems natural to extrapolate this to the case of an ideally uniform microscopic (or nanoscopic) region that would have zero hysteresis. In our picture hysteresis is the result of interaction between different phases in a multi-domain macroscopic sample, as will be detailed below.

At a given temperature T inside the major hysteretic loop, some parts of the film have $T_C(x, y) < T$ and some $T_C(x, y) > T$. In the first approximation, the boundary wall between the S and M phases is determined by the condition $T_C(x, y) = T$. In this approximation, the wall is highly irregular and its ruggedness corresponds to the scale at which one can define the local $T_C(x, y)$, *i.e.* the characteristic length scale of the nanoscopic phase domains. On closer inspection, however, we need a refinement that takes into account the *boundary energy* of the phase domain wall itself. This boundary energy is positive and to minimize its contribution to the free energy, the domain walls are relatively smooth.

Let us examine the process of boundary motion. For concreteness, let us consider the heating branch. Below the percolation transition, the M-phase resembles lakes in the S-phase mainland. With increasing temperature, the area of the M-phase increases, so the lakes grow in size. When a boundary of a given lake is far from the other lakes, infinitesimal $\pm dT$ changes boundary length by infinitesimal amount $\pm dL$, and the lake area by $\pm dA_M$. In other words, when the lakes are sufficiently separated, we envision a continuous, reversible, hysteresis-free process of M $\leftrightarrow$ S area redistribution, with neighboring configurations differing *microscopically,* what we call *area breathing.*

Let us now look at the formation of a link between two neighboring regions, which is the elementary step in the topological evolution of a global percolation picture. Consider two metallic lakes that are about to merge. Since the boundary is smooth, at some T the distance between the lakes becomes smaller than the radius of curvature of either lake at the point they will eventually touch. Therefore, at some critical T_{CR} the following two configurations will have equal energies: one comprising two disconnected M-phase lakes that are nearly but not quite touching, and the other with a finite link formed, Figs. 7(a) and 7(b) respectively.

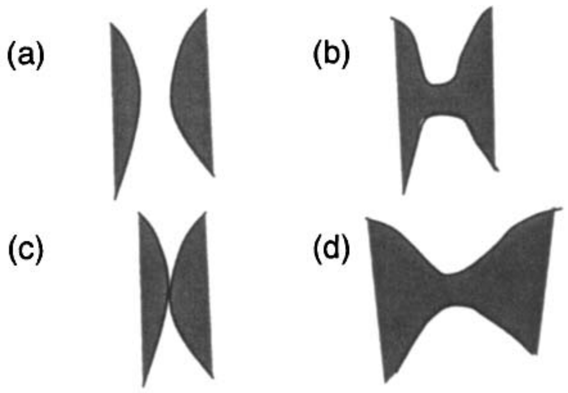


Figure 7. Semiconductor-metal boundary, with the metallic phase shown as shaded. Top row (a), (b) corresponds to temperature T_1 and the bottom row (c), (d) to temperature $T_2 > T_1$.

Both configurations are characterized by equal boundary lengths and therefore have equal free energy. In the thermodynamic sense one could call the T_{CR} the critical temperature for the link formation, if we could wait long enough. The actual transition forming a local link, however, does not occur at that temperature because of an immense *kinetic* barrier between these two *macroscopically* different configurations. The transition occurs at a higher $T_0 = T + \Delta T^*$ when it is actually forced, *i.e.* when the two phases touch at a point. Here ΔT^* is the *coercive* temperature. As can be seen, in our picture coercive temperature arises as a result of having a boundary between different phases; it does not preexist within each domain. We associate the steep slopes of the major loop with the quasi-continuous formation of such links, *i.e.* with local topological changes, specifically with the merger of metallic lakes on the HB and semiconductor lakes on the CB. Near T_S and T_M, the global map consists of widely separated M and S lakes, respectively. In these regions we expect to see nonhysteretic behavior; indeed, looking at Fig. 2 we see that $\Delta T = 10$ °C NHBs are largely non-hysteretic within ~20 °C from T_S and ~15°C from T_M.

Consider now a small excursion backwards from T_0 on the HB. As the temperature decreases, some of the M-phase recedes and the S-phase grows, changing the geometry of the global two-phase map. However, topologically, the last formed M-link does not disappear immediately for the same kinetic reason. One has two S regions that need to touch in order to wipe out the M-link. It takes a backward excursion of amplitude ΔT^* to establish an S-link and thus disconnect the last M-link. So long as we are within ΔT^*, *i.e.* stay on the same NHB, the area of S and M domains changes continuously, but the topology is stable and no new links are formed. Within the range of that stable or *frozen topology*, $\Delta T^* = \Delta T_{NHB}$, the resistivity of NHB will be single-valued and its temperature dependence will be controlled by the percolating semiconductor phase. This explains why NHBs have semiconducting slopes.

We introduced here a notion of *frozen topology*, which describes a global two-phase map with stable connecting links between regions of the same phase, while geometrical shape of the two phases is allowed to change, or breathe. One can also envision NHBs in which this breathing stops altogether, no switching between phases taking place, keeping the geometrical outlines fixed, this being the case of *frozen geometry*. Clearly *frozen geometry* implies *frozen topology*, but not the other way around. We will see below that both conditions actually exist and can be observed in a transitioning VO_2.

We can further address the observed phenomenon of TCR enhancement in some of the NHBs, where TCRs exceed the S-phase value – see Fig. 6. We note that for all NHBs, on both branches of the major loop, higher T implies increased fraction of M-phase, even if no new links are formed. This smooth change in geometry will produce additional temperature dependence adding to the semiconductor slope within a NHB, *i.e.* higher TCR. This effect will be stronger in the temperature intervals where this smooth change in geometry is more active, and it will cease to exist in the regions of *frozen geometry*, where the NHB slope will be determined only by the percolating S-phase. We will refer to this effect in explaining our optical reflectance measurements below.

Theoretical models that explain the data are good, but those that correctly predict new behavior are better. What can we predict based on our model?

We recall that hysteresis takes place in all physical properties that change in the phase transition. In addition to resistivity, optical constants, such as the refractive index n, change in the VO_2 transition. In a small uniform domain, n changes abruptly at T_C, from n_S to n_M. In a macroscopic film, there is a distribution of $T_C(x, y)$ and therefore, in a temperature range of co-existing S and M phases, $T_S < T < T_M$, the sample will contain some S domains with n_S and some M domains with n_M. As phases transform with changing T, the macroscopic film will exhibit a temperature-dependent hysteretic $n(T)$, much like the previously discussed $R(T)$. Indeed, optical properties exhibit hysteretic transition as a function of temperature. Among them, optical reflectivity is probably the easiest optical quantity to measure, although the details of how optical reflectivity behaves in a phase transition in a film with thickness comparable with the wavelength of light have to do not only with $n(T)$ but also with thin-film interference. This issue is discussed in detail in Ref. 8 (and in references therein), and we will not repeat the explanation here. The result is that, when T increases from 25 °C to 100 °C and decreases back to 25 °C, optical reflectivity $R_\lambda(T)$ at a fixed wavelength λ traces a hysteresis loop between certain values $R_{\lambda S}$ and $R_{\lambda M}$, these values being determined by a number of factors, including the choice of λ, film thickness, n_S and n_M. We can further expect that backward excursions from the points on the major optical loop will produce optical NHBs, for the same reason they exist in $R(T)$, and indeed we observed them, as will be seen below.

Based on our theoretical model, what specific behavior can we predict in optical reflectivity *vs.* T? Reflectivity is measured at wavelengths exceeding the nano-domain scale. It is largely determined by thin-film interference, with two-phase material between the film surfaces (in the Fabry-Perot resonator) having the average value of n, the latter depending on the volume fraction of S and M phases (volume fraction of n_S and n_M), which in a thin film becomes the area fraction of the two phases. In contrast, resistivity depends on the connectivity (links) of the percolating phase. Let A_S and A_M be the areas of S-phase and M-phase in a two-dimensional sample (thin film), so that the total sample area is $A = A_S + A_M$. Clearly, as A does not depend on T, $dA_M/dT = -dA_S/dT$, *i.e.* the area of one phase grows at the expense of the other. The optical slope dR_λ/dT observed in optical NHB must be proportional to this area redistribution slope, with the maximum in dR_λ/dT reflecting the maximum in dA_M/dT. According to our explanation of TCR enhancement over the S-phase value, the highest TCR should be found at the point that has the highest rate of area redistribution $dA_M/dT = -dA_S/dT$, and, therefore, at the point which corresponds to the maximum optical NHB slope dR_λ/dT. Likewise, zero rates of area redistribution (*frozen geometry*) corresponding to $dR_\lambda/dT = 0$, should correspond to unenhanced S-phase TCR. In the next section we will have a chance to verify these specific predictions.

The percolation picture also helps to understand why dR_λ/dT exhibits such a maximum in the first place. With changing temperature, the boundary moves, each section of the boundary line advancing in the direction normal to this line at any given temperature. It is clear that the highest rate of change of the area of each phase will therefore occur when the boundary is the longest, *i.e.* at the percolation transition. If our picture is correct, then the observed peak in TCR *vs.* T_0 in Fig. 6 occurs right at the percolation transition, allowing its detection. Finally, we see in Fig. 6 that above the peak, *i.e.* above the percolation transition, TCRs decrease quickly. Indeed, if the M-phase percolates, it shorts out the S-phase, and such decrease is to be expected.

6. NHBs in optical reflectivity – testing the theoretical model

In order to test some of the predictions of our model, simultaneously with the resistive measurements presented in Figs. 2 and 4, we measured reflectivity $R_\lambda(T)$ at a fixed wavelength λ. The optical signal was taken from the space between the voltage probes used in a four-probe resistivity measurement. In Fig. 8(a) we show major loop and minor hysteresis loops with $\Delta T = 10$ °C in optical reflectivity $R_\lambda(T)$ (*i.e.* this is an optical analog of Fig. 2). Some of the minor loops degenerate into optical NHBs, just like in resistivity measurements of Fig. 2. Additionally, we see that some minor loops are T-dependent, while others are not, and that T-dependent ones tend to be double-valued, while T-independent ones degenerate into NHBs.

In Fig. 8(b) we plot data for the same sample with shorter 5 °C excursions (i.e. this figure is an optical analog of Fig. 4). We see that all minor loops degenerated into NHBs, in complete analogy to Fig. 4, and according to our model of Section 5. Some of these NHBs, those closer to the center of the major loop, are T-dependent, while others closer to the loop limits are not. The T-dependent ones correspond to hysteretic minor loops in Fig. 8(a), while T-independent NHBs correlate with non-hysteretic behavior in Fig. 8(a).

Figure 9 combines the optical slopes dR_λ/dT of NHBs from Fig. 8(b) with the TCR data from Fig. 4 for both HB and CB branches. Clearly, the predictions of

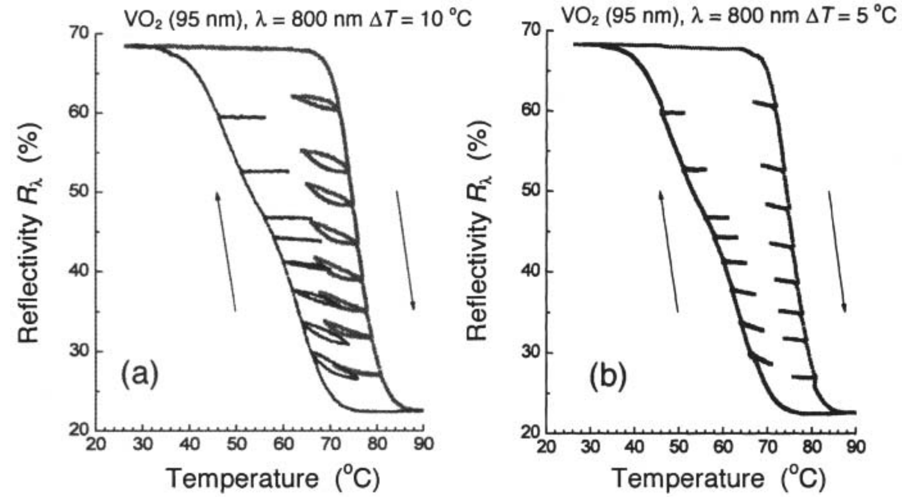


Figure 8. (a) Major hysteresis loops in optical reflectivity with minor loops, excursion length $\Delta T = 10$ °C ($\lambda = 800$ nm); (b) same data with $\Delta T = 5$ °C.

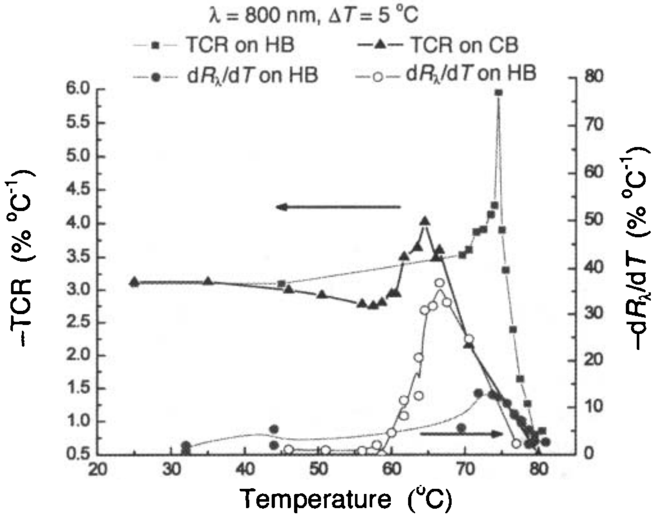


Figure 9. A combined plot of TCR and dR_λ/dT *vs.* T from various NHBs around the major loop plotted on one graph, exhibiting expected correlations.

our model are borne out: the peaks in dR_λ/dT on both branches coincide almost precisely with the peaks in TCR, and the region in which $dR_\lambda/dT \approx 0$ corresponds to TCR ≈ $(TCR)_S$. Further, if $dR_\lambda/dT \approx 0$ is a signature of a *frozen geometry*, it seems natural that, in the absence of S ↔ M transitions, minor loops should degenerate into NHBs more readily, as we observe in Fig. 8. The small (~2 degrees) misalignment of dR_λ/dT and TCR peaks in Fig. 9 is due to the fact that the factors that determine the overall shape of the major optical loop, such as interference effects,[8] should also to some extent affect the NHB slopes. Likewise, the relatively small deviation of the TCR in the $dR_\lambda/dT \approx 0$ region from the S-phase value – the observed shallow TCR minimum on the CB – can be ascribed to secondary causes. We do not think that these details disprove or diminish the impressive overall agreement of the main features predicted by our model.

The overall shape of the optical hysteresis and the range between pure-phase reflectivity values of $R_{\lambda S}$ and $R_{\lambda M}$ can be changed by making measurements at different wavelengths λ. In addition to the $\lambda = 800$ nm measurements shown in Fig. 8, we also measured $R_\lambda(T)$ at $\lambda = 503$ nm. The $\lambda = 503$ nm optical loop looks entirely different than the one shown in Fig. 8. Yet the dR_λ/dT exhibits the same peaks as in Fig. 9 and these peaks correlate with TCR peaks, confirming that we are indeed observing intrinsic S ↔ M area redistribution and not the effects of interference, *etc.*

Summing up, we proposed a qualitative theoretical picture, described in Section 5, without detailed mathematical modeling, which would be rather involved (perhaps a task for the future). We describe hysteresis in the framework of a percolation transition, as arising from interaction of S and M phases, specifically ascribing strong hysteretic effects to link formation between the two phases. This model corresponds to a number of observed features in both resistive and optical data. The most impressive agreement is achieved in the explanation of a subtle effect of TCR enhancement over the S-phase value in a range of NHBs where optical slopes correlate with TCRs, leaving no doubt that we understand the reason for TCR enhancement correctly.

8. Resistive microbolometers in NHB regime

A detailed discussion of the NHB method in the context of IR visualization with resistive microbolometers (the uncooled focal plane array or UFPA technology) is given in Ref. 1. Here we only give a brief summary. Good quality, single-phase VO_2 films may replace VO_X mixed oxide as sensor material in pixilated bolometric array. Despite using VO_2, hysteresis is eliminated when a sensor array operates within an NHB. The NHB will be chosen on the basis of its desired resistance, which can be adjusted over a wide range in order both to match the readout circuit amplifier and to minimize noise and Joule heating. The resistance will be 2–3 orders of magnitude smaller than the unacceptably high VO_2 resistance at 25 °C, while maintaining semiconducting TCR. We have seen NHBs with resistance as low as 20 Ω having TCR ≈ –4% $°C^{-1}$. We can also choose an NHB to maximize

TCR, which, as we have seen, varies between different NHBs around the major loop, peaking at the percolation transition at a value as high as 6% $°C^{-1}$. The operating temperature T_{OP} (*i.e.* the temperature at which the sensor array is stabilized awaiting the projected IR signal) will be chosen within an NHB, either near one of the ends or in the middle of the available range (total NHB width) ΔT_{NHB} = 4–5 °C. Because of the hysteresis, the process of reaching T_{OP} starting from room temperature requires performing specific heating and cooling steps. Specifically, positioning an array at T_{OP} will require: on the HB, warming up to T_0 and cooling down to T_{OP}; on the CB, warming up to above T_M, cooling down to T_0, and again warming up to T_{OP}. If T_{OP} is chosen in the middle of an NHB, the last step requires cooling down from T_0 to $(T_0 - \Delta T_{NHB}/2)$ on the HB or warming up from T_0 to $(T_0 - \Delta T_{NHB}/2)$ on the CB.

9. Summary

We found a regime of operation within a hysteretic phase transition that avoids hysteresis within a limited range of a transition-controlling parameter. This may benefit applications utilizing a phase transition but suffering from complications introduced by hysteresis. One such application is IR visualization with resistive microbolometers. We propose to use VO_2 as a sensor material, operating it in the nonhysteretic branch (NHB) regime. Partial shorting out of the S-phase by the M-phase lowers sensor resistance; the degree of admixture of M-phase determines the value of this resistance, making it adjustable. At the same time, because of the S-phase percolation, NHB has semiconducting $R(T)$. The TCR in this case can be considerably higher than in the pure S-phase due to a smooth change of geometry (but not the topology) of S-M areas within an NHB. A similar NHB regime exists in optical reflectivity. Optical reflectivity slopes correlate with TCRs, revealing the reason for TCR enhancement in some of the NHBs over the S-phase value. We believe that NHB regime may be implemented in any hysteretic phase transition, removing problems associated with hysteresis and potentially benefiting electronic applications.

Acknowledgments

We thank Dr. David Westerfeld for developing Labview-based automated resistivity measurement system and Dr. Arsen Subashiev for useful theoretical discussions. This work was supported by the New York State Foundation for Science, Technology and Innovation (NYSTAR) through the facilities of the Center for Advanced Sensor Technology at Stony Brook.

References

1. M. Gurvitch, S. Luryi, A. Polyakov, and A. Shabalov, "Non-hysteretic behavior inside the hysteresis loop of VO_2 and its possible application in infrared imaging," *J. Appl. Phys.* **106**, 104504 (2009).
2. F. J. Morin, "Oxides that show a metal-to-insulator transition at the Neel temperature," *Phys. Rev. Lett.* **3**, 34 (1959); N. Mott, *Metal-Insulator Transitions*, London: Taylor & Francis, 1997.
3. C. N. Bergland and H. J. Guggenheim, "Electronic properties of VO_2 near the semiconductor-metal transition," *Phys. Rev.* **185**, 1022 (1969).
4. R. S. Scott and G. E. Frederics, "Model for infrared detection by a metal-semiconductor phase transition," *Infrared Phys.* **16**, 619 (1976).
5. V. Yu. Zerov, Yu. V. Kulikov, V. N. Leonov, V. G. Malyarov, I. A. Khrebtov, and I. I. Shaganov, "Features of the operation of a bolometer based on a vanadium dioxide film in a temperature interval that includes a phase transition," *J. Opt. Technol.* **66**, 387 (1999).
6. B. E. Cole, R. E. Higashi, and R. A. Wood, "Monolithic two-dimensional arrays of micromachined microstructures for infrared applications," *Proc. IEEE* **86**, 1679 (1998).
7. T. G. Lanskaya, I. A. Merkulov, and F. A. Chudnovskii, "Hysteresis effects at the semiconductor-metal phase transition in vanadium oxides," *Sov. Phys. Solid State* **20**, 193 (1978), and a number of more recent papers treating hysteresis in VO_2.
8. M. Gurvitch, S. Luryi, A. Polyakov, *et al.*, "VO_2 films with strong semiconductor to metal phase transition prepared by the precursor oxidation process," *J. Appl. Phys.* **102**, 033504 (2007).

This page intentionally left blank

Exploratory Studies on Silicon-Based Oxide Fuel Cell Power Sources Incorporating Ultrathin Nanostructured Platinum and Cerium Oxide Films as Anode Components

Bo-Kuai Lai, A. C. Johnson, H. Xiong, C. Ko, and S. Ramanathan
School of Engineering and Applied Sciences, Harvard University
Cambridge, MA 02138, U.S.A.

1. Introduction

Solid oxide fuel cells (SOFCs) convert chemical energy from hydrogen or hydrocarbons into electrical energy with high power density, high energy efficiency, and low carbon footprints. Their potential applications could cover the span from large-scale power plants to portable electronics.[1,2] SOFCs that are of small form factor and can be manufactured at low cost have increasingly attracted research attention.[2-10] Such SOFCs are usually referred to as micro (μ) or thin-film (TF) SOFCs because of their small form factor and/or the microfabrication and thin film deposition techniques employed.[8,11] In addition, TFSOFCs may also serve as model systems to elucidate the roles of thickness, microstructure, surfaces, and interfaces in further enhancing performance of intermediate- and low-temperature SOFCs.[12-18]

The primary physical structure of a TFSOFC is a membrane consisting at least three layers of sub-μm thin films – namely, cathode, electrolyte, and anode. During SOFC operation, oxygen molecules are reduced to oxygen ions on cathode surface or triple phase boundaries (TPBs.) The oxygen ions are then transported (via ion-vacancy exchange) through electrolyte and anode to combine with hydrogen on anode surface or at TPBs. The charge flow of oxygen ions from cathode to anode is accompanied with a current flow of electrons from anode to cathode through the external circuit. The overall performance of the heterostructures is extremely sensitive to properties of each layer and interfaces between layers.

In the past decade, development of TFSOFCs[3] has evolved from microfabrication of thermomechanically stable yttria-stabilized zirconia (YSZ) electrolyte membranes,[19,20] to implementation of Pt and Pt-based composite cathodes,[10,21,22] to the replacement of Pt or Pt-based cathodes with oxide cathodes,[2,7,9,11,22,23] as well as performance enhancement by inserting an interlayer between YSZ electrolytes and Pt cathodes or using corrugated membranes.[10,24] To date, because of its excellent electrochemical catalytic properties and the lack of satisfactory anode alternatives, pure Pt is still used as the anode for majority of the TFSOFCs reported in the literature.[2,7,9,22,24] One specific issue with pure Pt anodes

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

in TFSOFCs is the instability of porous microstructures. In order to increase area of TPBs, at which oxygen reduction occurs, Pt is usually deposited at high pressure and annealed at elevated temperatures to yield a porous microstructure. However, as reported in Refs. 9, 10, and 21, such porous microstructures may not be stable – namely, Pt thin films undergo grain coarsening and then dewetting with increasing temperature. The dewetting effect is due to the capillary instability associated with high surface-to-volume ratioa and has been observed for other ultrathin metal films.[25-27] Its occurrence is usually preceded by formation of pores[21] in otherwise continuous thin films[28,29] and then proceeds until the films agglomerate to form isolated islands.[9] As a consequence, the discontinuous Pt thin films become laterally insulating.

Besides further enhancing microstructure stability of Pt thin films, a more critical challenge for TFSOFCs is using an oxide or oxide-based anode to replace Pt or to reduce volume of the Pt used,[21] which is important for cost reduction, as well as scientific perspectives. The goal of this article is to report our recent progress in developing microstructurally stable Pt thin films and exploring functional oxides on the anode side of TFSOFCs. Since the thin films could cover electrolyte, Si (or Forturan), and/or Si_3N_4 surfaces in current fabrication processing of TFSOFCs,[2,9,10] it is important to understand behavior of oxide thin films on different surfaces because their properties might be substrate-dependent. For instance, Pt reacts with Si in 200−450 °C range but not with YSZ.[30-32] Here, we will focus our efforts on oxide thin films deposited on YSZ and Si_3N_4 surfaces and characterize their properties that are of direct relevance to TFSOFC performance. The properties investigated include: microstructure, crystallization behavior, film stresses, and electrical conductivity at room temperature (RT). We have also fabricated TFSOFCs based on the thin films studied, while using YSZ as electrolyte and Pt or lanthanum strontium cobalt ferrite (LSCF) as cathodes, to investigate their functionality on the anode side of TFSOFCs.

Unlike metal deposition, oxide materials are relatively difficult to sputter-deposit in a porous form. Therefore, oxides that are suitable for TFSOFC anode applications need to be good ionic and adequate electronic conductors. The former allows oxygen ions from the electrolyte to reach the anode surface via ion-vacancy exchange, while the latter enables electron collection and delivery to the external circuits.[33] The oxides we chose to investigate were cerium oxides (ceria, CeO_x), despite the fact that undoped ceria has heretofore not been recognized as a promising anode material.[34] The advantages of ceria include good mixed conduction and enhanced electronic conduction in nanostructured forms.[35,38-40] Enhanced performance of TFSOFCs has been reported when using Gd-doped ceria (GDC) interlayer between the Pt cathode and YSZ electrolyte,[10] but it is still unclear if such enhancement can be observed when using GDC or ceria as the anode interlayer.[41] Moreover, the mechanisms (ionic conductivity, electronic conductivity and/or reduction activity at the GDC/Pt/air TPBs) responsible for the enhancement are still not well understood. Also, ceria is widely used as matrix or active components in many catalytic composites. Since electrochemical functionality and thermomechanical stability are complex and intertwined phenomena for the anodes in TFSOFCs, investigation of undoped ceria could provide valuable insight and guidance for further electrode research using doped ceria or ceria-based composites.

In this report, we show that microstructural stability of porous Pt thin films can be enhanced by simply increasing its thickness while functional fuel cell performance is still maintained. We also examine the suitability of ceria for the anode side of TFSOFCs. Our results indicate that, when using oxides in TFSOFCs, it is critical to understand temperature-, substrate-, and environment-dependence of phase transition, microstructure change, interfacial reaction, and stress evolution in order to maintain stability and electrochemical functionality of TFSOFC membranes. We anticipate that the results presented here can provide insights and guidance for further anode development using complex oxides or composites for intermediate- or low-temperature SOFCs.

2. Experimental

Porous platinum (Pt) was sputtered at a dc power of 250 W and a pressure of 75 mTorr. Cerium oxides were deposited using a pure cerium (Ce) target without substrate heating. The dc power for the cerium target was 100 W and the process gas, 10 sccm of Ar and 10 sccm of O_2, was maintained at 5 mTorr pressure during sputtering. A custom $La_{0.6}Sr_{0.4}Co_{0.8}Fe_{0.2}O_3$ (LSCF) target and an 8% yttria-doped zirconia (YSZ) target operated at a gun power of 60 W and 100 W, respectively, were used to deposit LSCF and YSZ thin films. Thickness of LSCF thin films was controlled to be in 15–20 nm range to avoid microcracks while maintaining adequate electrical conductivity. Deposition and optimization of LSCF thin films for TFSOFCs applications are described in detail elsewhere.[11,23] We used single crystalline YSZ (100) and Si_3N_4-coated Si wafers as substrates.

Unless otherwise specified, annealing was performed in air for 2 hours at ramping and cooling rates of 5 °C/min in a Thermolyne 21100 tube furnace. Grazing incidence x-ray diffraction (XRD) was performed with a Scintag 2000 diffractometer using Cu K_α radiation at an incidence angle of 1°. Microstructures were investigated by Carl Zeiss Ultra 55 field emission scanning electron microscopy (SEM). Temperature-dependent film stresses were measured by a Toho FLX-2320-S thin film stress measurement system with an identical temperature profile as the annealing experiments. The system measures stress-induced curvature change of 4" substrate wafers prior to and after thin film deposition, with the film stresses determined from Stoney equation.[42] Nominal substrate thicknesses of 525 μm and biaxial modulus of 181 GPa for Si were used in the calculation. Lateral electrical conductivity at RT was measured by a Creative Design Engineering ResMap 168 four-point probe resistivity mapping system. TFSOFCs were fabricated by photolithography similar to the procedure described elsewhere and measured using the same apparatus.[9]

3. Results and discussion

- *Pt thin film anode*

We deposited Pt thin films with a nominal thickness of 130 nm. This value is thicker than the Pt thin films that suffer dewetting-induced microstructure instability.[9,21] Figures 1(a)–(e) show SEM images of morphology evolution of 130 nm Pt/YSZ thin films annealed at different temperatures. As seen in Fig. 1(a), because of high deposition pressure, Pt particulates were loosely packed on surface. Upon annealing at 300 °C, size of the particulates slightly increases. A distinct morphology change – namely, coarsening of the particulates – was observed when the films were annealed at 400 °C, as shown in Fig. 1(c), and that continues with further annealing to 500 °C. Upon annealing at 600 °C, gaps between coarsened particulates become narrower and the films form a more continuous and smooth morphology – in other words, the films begin to slightly wet the surface. Similar wetting phenomenon was also observed for the 500 °C-annealed thin film with additional 24 hours of annealing and for thin films annealed in the SOFC operation environment[9] – *i.e.* a flowing gas mixture (5% H_2 and 95% Ar) at a rate of 25 sccm, see Fig. 1(f).

Effect of annealing temperature on as-deposited Pt/YSZ thin films was further investigated by using the four-point probe measurement. As seen in Fig. 2, the measured electrical conductivity corresponds well to the morphology evolution displayed in Fig. 1. The as-deposited and 300 °C-annealed thin films that exhibit similar morphology possess similar electrical conductivity. When the Pt particulates start to coarsen above 300 °C, electrical conductivity also begins to increase rapidly with increasing temperature. The observed wetting in the 600 °C-annealed, 500 °C-24 hour-annealed, 550 °C-H_2-annealed, and 600°C-24 hour-annealed samples leads to electrical conductivity saturation to values that are ~60% of the bulk conductivity ($\sim 96.6 \times 10^3$ S/cm). The results, shown in Figs. 1(e)–(f) and 2(a), also indicate that oxidizing or reducing atmosphere during annealing has little effect on the properties of Pt/YSZ thin films.

Compared to similar Pt thin films that were sputtered on Si substrates at high deposition pressure in Ref. 21, our results show that Pt/YSZ thin films also undergo morphology change from particulate coarsening to surface wetting upon annealing. However, the morphology of our Pt/YSZ thin films appears relatively stable when annealed in 500–600 °C range and no sign of dewetting was observed. The difference in morphology at 600 °C is mainly due to the fact that the Pt films in our study are thicker. Different interfacial energy of Pt/Si and Pt/YSZ interfaces and absence of interfacial reactions in the Pt/YSZ system in this temperature range might also be partially responsible for the observations. The result demonstrates that, for Pt/YSZ thin films, it is possible to enhance morphology stability, at the cost of less porosity, by simply increasing film thickness.

One of the central concerns for TFSOFCs is the stress-induced thermomechanical instability in ultra-thin film membranes.[19,20] Since morphology evolution at temperatures below 500 °C is similar for both Pt/Si_3N_4 and Pt/YSZ systems, Pt thin films were deposited onto Si_3N_4-coated 4" Si wafers using identical deposition parameters as the Pt/YSZ thin films to investigate the temperature-related stress evolution below 500 °C. Fig. 2(b) shows stress evolution of such Pt/Si_3N_4 thin films. Low film stresses at RT are attributed to the loosely packed Pt particulates. Except for the stress deviation that occurs in 300–420 °C range upon heating, film stresses decrease with increasing temperature due to larger thermal expansion of Pt than Si. The tensile stresses are due to the pulling force between adjacent particulates during coarsening.[42]

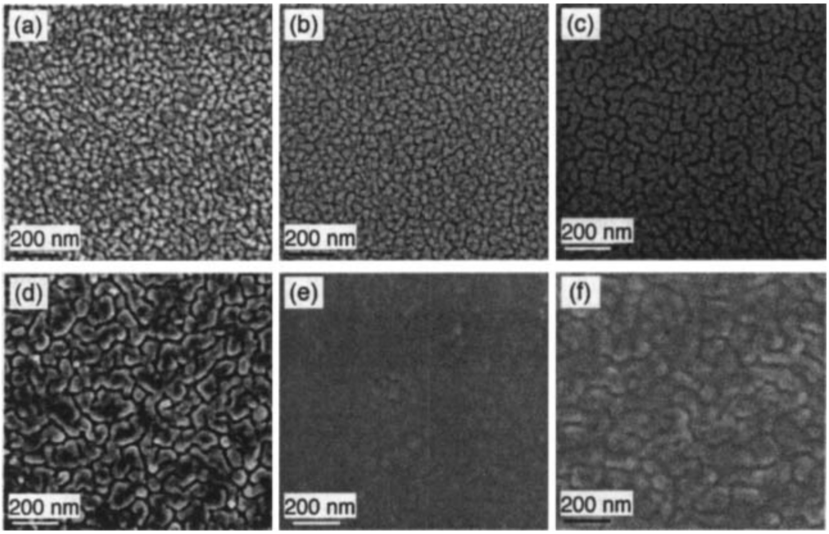


Figure 1. SEM images of Pt/YSZ thin films that were (a) as-deposited without substrate heating, annealed in air at (b) 300 °C, (c) 400 °C, (d) 500 °C, and (e) 600 °C, and (f) annealed in flowing 5% H_2 at 550 °C.

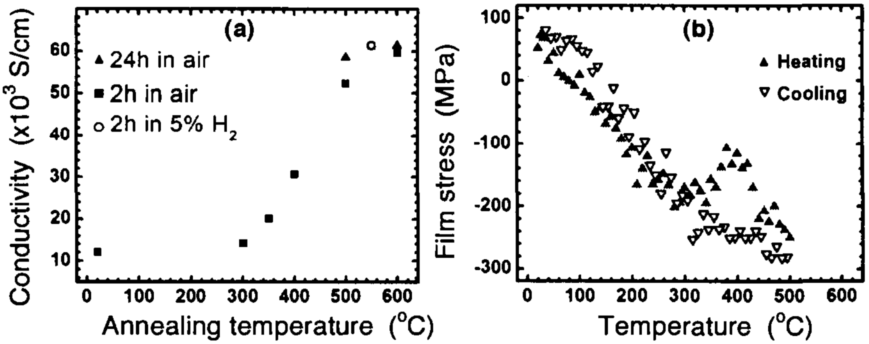


Figure 2. (a) RT electrical conductivity of Pt/YSZ thin films annealed at various temperatures for 2 hours in air (■), 24 hours in air (▲), and 2 hours in flowing 5% H_2 (o). (b) temperature dependence of film stresses in 130 nm Pt/Si_3N_4 thin films upon heating (▲) and cooling (▽).

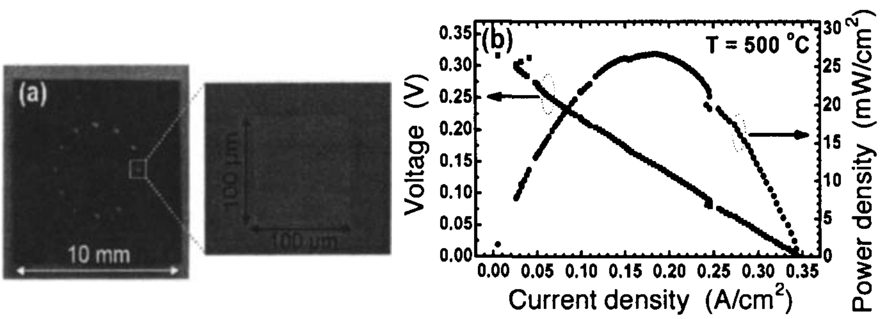


Figure 3. (a) Micrographs of TFSOFC array with twelve 100x100 µm square membranes on a 10×10 mm Si chip and a close up micrograph for one of the membranes; (b) power density as a function of current density measured at 500 °C.

Performance of TFSOFC array based on 20 nm-LSCF/60 nm-YSZ/130 nm-Pt membranes (LYP, hereafter) was measured by placing a probe at the center of the array. Figure 3(a) shows a TFSOFC chip that contains twelve 100×100 µm square membranes in a circular arrangement. The low open-circuit voltage (OCV) of ~0.32 V of the TFSOFC array is primarily attributed to the probing method we used – namely, the overall OCV is determined by the TFSOFC with the lowest OCV among the twelve. As pointed out in Ref. 22, this probing method could also lead to lower OCV because the 20 nm LSCF cathode might not be conductive enough as the current collector.[11] Nevertheless, as seen in Fig. 3(b), our measurements at 500 °C shows an average power density of ~27 mW/cm² per membrane. Despite the low OCV and insufficient current collection, the power density is only a factor of two lower than our previous measurements on a single TFSOFC using a thinner Pt anode but thicker YSZ electrolyte.[9] Therefore, the current results indicate a balance of microstructure stability and TFSOFC performance is achievable by carefully controlling thickness of porous Pt thin films.

- *Ceria thin films*

We deposited 30 nm CeO_x concurrently onto Si_3N_4 and YSZ substrates, referred to as CeO_x/Si_3N_4 and CeO_x/YSZ thin films, respectively. Figure 4(a) shows XRD patterns of as-deposited and 500 °C-annealed CeO_x/Si_3N_4 thin films, with the corresponding SEM images displayed in Figs. 5(a) and (b), respectively. The as-deposited CeO_x/Si_3N_4 thin films exhibit crystalline CeO_2 phases (JCPDS No. 03-065-2975) and a smooth surface morphology. After 500 °C annealing, no change in the phase and morphology was observed except for an angular shift of XRD peaks. The absence of morphology change in the presence of apparent lattice change was also observed in nanocrystalline $CeO_{2-\delta}$ on Si and on sapphire.[43,44]

The shift of XRD peaks towards a higher 2-theta indicates a decrease in the lattice constant. For $CeO_{2-\delta}$, its lattice constant increases linearly with increasing oxygen deficiency because the formation of oxygen vacancies is charge-compensated by a $Ce^{4+} \rightarrow Ce^{3+}$ reduction and the ionic radius of Ce^{3+} is higher than of Ce^{4+}.[43, 45] Since the $CeO_{2-\delta}$ thin films were deposited in an oxygen-deficient vacuum environment[46] and concentration of oxygen vacancies is inherently higher[38] in grain boundaries, vacuum-deposited nanocrystalline $CeO_{2-\delta}$ is more deficient in oxygen and hence has a higher lattice constant than its bulk counterpart.[47] When annealed in air, oxygen incorporation occurs and leads to a decrease of the lattice constant. Similar phenomena have also been observed for as-deposited YSZ and LSCF thin films.[9,11]

As seen in the stress evolution plot of Fig. 4(b), the oxygen incorporation likely occurs at ~120 °C and induces a change in the sign of stress/temperature slope from negative to positive. These results are consistent with previous reports on oxygen storage and release at low temperatures.[45] However, although thermal expansion of CeO_2 is greater than Si, the slope of stress evolution remains positive above 120 °C upon heating. Upon cooling, the stress evolution also does not obey thermal expansion difference in 500–340 °C and 235–155 °C temperature ranges. These results resemble those reported in Ref. 43, for which an unusual lattice parameter change that is independent of thermal expansion difference was observed in nanocrystalline $CeO_{2-\delta}/Si$ thin films annealed in 150–450 °C range. Due to the excellent crystal structure and lattice constant matching between ceria and YSZ, the sputtered CeO_x thin film grows on YSZ (100) substrates in a highly textured manner, as indicated by the sole presence of (200) peak in Fig. 4(c).

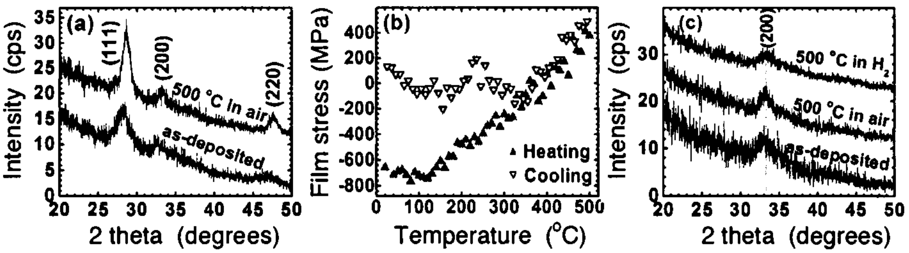


Figure 4. (a) XRD patterns of as-deposited and 500°C-annealed 30 nm CeO_x/Si_3N_4 thin films; (b) film stress as a function of temperature for CeO_x/Si_3N_4 films; (c) XRD patterns of 30 nm CeO_x/YSZ thin films that were as-deposited without substrate heating, annealed at 500 °C in air, and annealed at 500 °C in flowing 5% H_2.

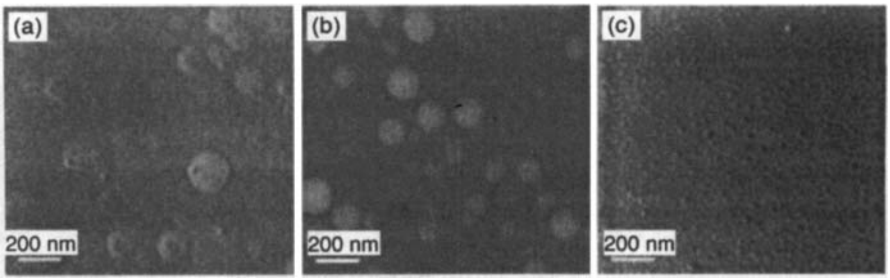


Figure 5. SEM images of 30 nm CeO_x/Si_3N_4 thin films: (a) as-deposited without substrate heating and (b) annealed at 500 °C in air. (c) 30 nm CeO_x/YSZ thin films that were annealed at 500 °C in flowing 5% H_2.

Compared to the CeO_x/Si_3N_4 thin films, the CeO_x/YSZ thin films exhibit a higher degree of structural and phase stability. Figure 4(c) shows that, after annealing in air at 500 °C, the (200) peak shifts only slightly. Moreover, CeO_x/YSZ thin films are not affected by the annealing environment. As seen in Fig. 4(c), the XRD patterns are essentially the same and no distinct morphology changes are observed when comparing the CeO_x/YSZ films that were annealed at 500 °C in air and in flowing 5% H_2 – see Fig. 5(c).

We fabricated TFSOFCs with isolated 100×100 μm 80nm-Pt/75nm-YSZ/30nm-CeO_x (PYC) and isolated 20nm-LSCF/75nm-YSZ/100nm-CeO_x (LYC) membranes.[9] Their OCV is smaller than 0.2 V at 500 °C and maximum power density (P_{MAX}) does not change when switching off H_2 flow. This result suggests that sputtered nanocrystalline CeO_x alone is inadequate for an anode, likely due to its low electrical conductivity and/or insufficient oxidation activity on ceria surfaces.[36, 38,48] The role of ceria as the interlayer was investigated by depositing Pt onto the anode side of the PYC and LYC membranes. Figure 6 shows P_{MAX} and OCV of 80nm-Pt/75nm-YSZ/100nm-CeO_x/80nm-Pt (PYCP) and 20nm-LSCF/75nm-YSZ/30nm-CeO_x/80nm-Pt (LYCP) membranes. Evidently, the PYCP membrane maintains an adequate OCV (> 0.9 V) in 300−500 °C range and exhibits a P_{MAX} of ~15 mW/cm^2 at 500 °C; while for the LYCP membrane, P_{MAX} ~ 3.6 mW/cm^2 at 500 °C. These results indicate that ceria can be used as the interlayer. Compared to our previously reported TFSOFCs of the same cathode materials,[9] both membranes exhibit similar temperature-dependent behavior of P_{MAX} and OCV. The thicker electrolyte interlayer in present study might result in lower ionic conductance, which is proportional to the product of ionic conductivity and inverse film thickness, and is likely responsible for the lower performance. We expect performance of the TFSOFCs can be improved by using a thinner electrolyte-interlayer and using GDC, which has a higher ionic conductivity than ceria and YSZ, as the interlayer. Since CeO_x/YSZ thin films exhibit excellent microstructure stability below 500 °C, ceria is suitable as the matrix of oxide-metal composites

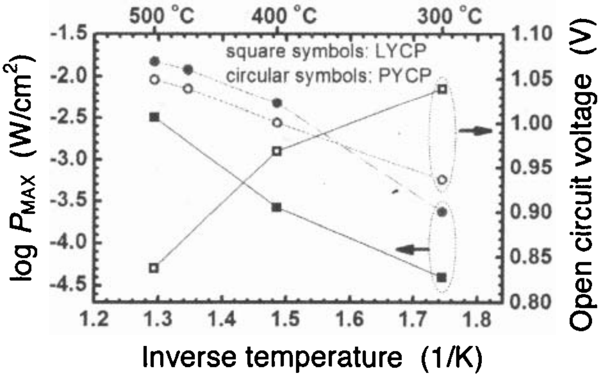


Figure 6. Peak power density P_{MAX} and open circuit voltage *vs.* temperature for PYCP (circular symbols) and LYCP (square symbols) TFSOFCs.

(*e.g.* by co-sputtering metal films and creating percolating microstructures), for intermediate- and low-temperature TFSOFCs. If adequate electronic conductivity can be achieved, such ceria-based composites may become promising for anode applications in TFSOFCs.

4. Conclusions

Platinum and ceria have been investigated systematically for anode applications in TFSOFCs. We found that microstructure stability of porous Pt thin films can be enhanced by increasing its thickness to minimize the dewetting effects. Fuel cell performance (~27 mW/cm^2 at 500 °C) is still maintained for TFSOFCs using such thicker Pt anodes, with ultra-thin LSCF and YSZ as the cathode and electrolyte, respectively. Sputtered CeO_x films possess excellent microstructure and phase stability across the relevant temperature range for both reducing and oxidizing environments, and are compatible with YSZ and Pt films on the anode side. TFSOFCs incorporating ceria thin films on the anode side have been fabricated and the performance of Pt/YSZ/CeO_x/Pt and LSCF/YSZ/CeO_x/Pt membranes has been successfully demonstrated to be 15.0 and 3.6 mW/cm^2 at 500 °C, respectively. We anticipate these results to be of relevance to advancing the science and technological applications of thin-film solid oxide fuel cells.

Acknowledgments

The work is supported by SiEnergy Systems, the Harvard University Center for the Environment and Harvard SEAS.

References

1. S. C. Singhal and K. Kendall, *High Temperature Solid Oxide Fuel Cells: Fundamentals, Design and Applications*, Oxford, UK: Elsevier (2003).
2. A. Bieberle-Hutter, D. Beckel, A. Infortuna, *et al.*, "A micro-solid oxide fuel cell system as battery replacement," *J. Power Sources* **177**, 123 (2008).
3. A. F. Jankowski, J. P. Hayes, R. T. Graff, and J. D. Morse, "Microfabricated thin-film fuel cells for portable power applications," MRS Symp. Proc. **730**, V4.2.1 (2002).
4. D. Beckel, A. Bieberle-Hutter, A. Harvey, *et al.*, "Thin films for micro solid oxide fuel cells," *J. Power Sources* **173**, 325 (2007).
5. J. H. Joo and G. M. Choi, "Simple fabrication of micro-solid oxide fuel cell supported on metal substrate," *J. Power Sources* **182**, 589 (2008).

6. U. P. Muecke, D. Beckel, A. Bernard, *et al.*, "Micro solid oxide fuel cells on glass ceramic substrates," *Adv. Funct. Mater.* **18**, 3158 (2008).
7. S. Rey-Mermet and P. Muralt, "Materials and design study for micromachined solid oxide fuel cells membranes," *MRS. Symp. Proc.* **972**, AA07 (2007).
8. S. J. Litzelman, J. L. Hertz, W. Jung, and H. L. Tuller, "Opportunities and challenges in materials development for thin film solid oxide fuel cells," *Fuel Cells* **8**, 294 (2008).
9. A. C. Johnson, B.-K. Lai, H. Xiong, and S. Ramanathan, "An experimental investigation into micro-fabricated solid oxide fuel cells with ultra-thin $La_{0.6}Sr_{0.4}Co_{0.8}Fe_{0.2}O_3$ cathodes and yttria-doped zirconia electrolyte films," *J. Power Sources* **186**, 252 (2009).
10. H. Huang, M. Nakamura, P. C. Su, R. Fasching, Y. Saito, and F. B. Prinz, "High-performance ultrathin solid oxide fuel cells for low-temperature operation," *J. Electrochem. Soc.* **154**, B20 (2007).
11. B. K. Lai, H. Xiong, M. Tsuchiya, A. C. Johnson, and S. Ramanathan, "Microstructure and microfabrication considerations for self-supported on-chip ultra-thin micro-solid oxide fuel cell membranes," *Fuel Cells* **9**, 699 (2009).
12. T. Suzuki, Z. Hasan, Y. Funahashi, T. Yamaguchi, Y. Fujishiro, and M. Awano, "Impact of anode microstructure on solid oxide fuel cells," *Science* **325**, 852 (2009).
13. W. T. Bao, H. M. Guan, and J. H. Cheng, "A new anode material for intermediate solid oxide fuel cells," *J. Power Sources* **175**, 232 (2008).
14. D. J. L. Brett, A. Atkinson, N. P. Brandon, and S. J. Skinner, "Intermediate temperature solid oxide fuel cells," *Chem. Soc. Rev.* **37**, 1568 (2008).
15. A. Ignatiev, X. Chen, N. J. Wu, Z. G. Lu, and L. Smith, "Nanostructured thin solid oxide fuel cells with high power density," *Dalton Trans.* **40**, 5501 (2008).
16. H. Matsumoto, I. Nomura, S. Okada, and T. Ishihara, "Intermediate-temperature solid oxide fuel cells using perovskite-type oxide based on barium cerate," *Solid State Ionics* **179**, 1486 (2008).
17. J. M. Serra, V. B. Vert, O. Buchler, W. A. Meulenberg, and H. P. Buchkremer, "IT-SOFC supported on mixed oxygen ionic-electronic conducting composites," *Chem. Mater.* **20**, 3867 (2008).
18. H. Xiong, B.-K. Lai, A. C. Johnson, and S. Ramanathan, "Low-temperature electrochemical characterization of dense ultra-thin lanthanum strontium cobalt ferrite ($La_{0.6}Sr_{0.4}Co_{0.8}Fe_{0.2}O_3$) cathodes synthesized by rf-sputtering on nanoporous alumina-supported Y-doped zirconia membranes," *J. Power Sources* **193**, 589 (2009).
19. V. T. Srikar, K. T. Turner, T. Y. A. Ie, and S. M. Spearing, "Structural design considerations for micromachined solid-oxide fuel cells," *J. Power Sources* **125**, 62 (2004).
20. Y. H. Tang, K. Stanley, J. Wu, D. Ghosh, and J. J. Zhang, "Design consideration of micro thin film solid-oxide fuel cells," *J. Micromech. Microeng.* **15**, S185 (2005).

21. X. H. Wang, H. Huang, T. Holme, X. Tian, and F. B. Prinz, "Thermal stabilities of nanoporous metallic electrodes at elevated temperatures," *J. Power Sources* **175**, 75 (2008).
22. X. R. Jiang, H. Huang, F. B. Prinz, and S. F. Bent, "Application of atomic layer deposition of platinum to solid oxide fuel cells," *Chem. Mater.* **20**, 3897 (2008).
23. B.-K. Lai, A. C. Johnson, H. Xiong, and S. Ramanathan, "Ultra-thin nanocrystalline lanthanum strontium cobalt ferrite ($La_{0.6}Sr_{0.4}Co_{0.8}Fe_{0.2}O_{3-\delta}$) films synthesis by rf-sputtering and temperature-dependent conductivity studies," *J. Power Sources* **186**, 115 (2009).
24. P. C. Su, C. C. Chao, J. H. Shim, R. Fasching, and F. B. Prinz, "Solid oxide fuel cell with corrugated thin film electrolyte," *Nano Lett.* **8**, 2289 (2008).
25. A. L. Giermann and C. V. Thompson, "Solid-state dewetting for ordered arrays of crystallographically oriented metal particles," *Appl. Phys. Lett.* **86**, 121903 (2005).
26. E. Jiran and C. V. Thompson, "Capillary instabilities in thin films," *J. Electronic Mater.* **19**, 1153 (1990).
27. K. Thürmer, E. D. Williams, and J. E. Reutt-Robey, "Dewetting dynamics of ultrathin silver films on Si(111)," *Phys. Rev. B* **68**, 155423 (2003).
28. X. Hu, D. G. Cahill, and R. S. Averback, "Dewetting and nanopattern formation of thin Pt films on SiO_2 induced by ion beam irradiation," *J. Appl. Phys.* **89**, 7777 (2001).
29. J. Petersen and S. G. Mayr, "Dewetting of Ni and NiAg solid thin films and formation of nanowires on ripple patterned substrates," *J. Appl. Phys.* **103**, 023520 (2008).
30. C. A. Crider, J. M. Poate, J. E. Rowe, and T. T. Sheng, "Platinum silicide formation under ultrahigh vacuum and controlled impurity ambients," *J. Appl. Phys.* **52**, 2860 (1981).
31. B. K. Lee, Y. H. Yu, B. S. So, *et al.*, "Electrical characterization of the platinum/YSZ interfaces in SOFCs via micro-contact impedance spectroscopy," *J. Electroceram.* **17**, 735 (2006).
32. A. Jaccoud, G. Foti, R. Wuthrich, H. Jotterand, and C. Comninellis, "Effect of microstructure on the electrochemical behavior of Pt/YSZ electrodes," *Topics Catalysis* **44**, 409 (2007).
33. J. B. Goodenough and Y. H. Huang, "Alternative anode materials for solid oxide fuel cells," *J. Power Sources* **173**, 1 (2007).
34. Q. X. Fu and F. Tietz, "Ceramic-based anode materials for improved redox cycling of solid oxide fuel cells," *Fuel Cells* **8**, 283 (2008).
35. S. Omar, E. D. Wachsman, and J. C. Nino, "Higher ionic conductive ceria-based electrolytes for solid oxide fuel cells," *Appl. Phys. Lett.* **91**, 144106 (2007).
36. J. L. M. Rupp and L. J. Gauckler, "Microstructures and electrical conductivity of nanocrystalline ceria-based thin films," *Solid State Ionics* **177**, 2513 (2006).

37. M. V. Ganduglia-Pirovano, A. Hofmann, and J. Sauer, "Oxygen vacancies in transition metal and rare earth oxides: Current state of understanding and remaining challenges," *Surf. Sci. Rep.* **62**, 219 (2007).
38. Y. M. Chiang, E. B. Lavik, I. Kosacki, H. L. Tuller, and J. Y. Ying, "Nonstoichiometry and electrical conductivity of nanocrystalline CeO_{2-x}," *J. Electroceram.* **1**, 7 (1997).
39. S. Kim and J. Maier, "On the conductivity mechanism of nanocrystalline ceria," *J. Electrochem. Soc.* **149**, J73 (2002).
40. X. Guo and R. Waser, "Electrical properties of the grain boundaries of oxygen ion conductors: Acceptor-doped zirconia and ceria," *Prog. Mater. Sci.* **51**, 151 (2006).
41. T. Omata, Y. Goto, and S. Otsuka-Yao-Matsuo, "Nanocrystals of zirconia- and ceria-based solid electrolytes: Syntheses and properties," *Sci. Technol. Adv. Mater.* **8**, 524 (2007).
42. F. Spaepen, "Interfaces and stresses in thin films," *Acta Mater.* **48**, 31 (2000).
43. A. Kossoy, Y. Feldman, E. Wachtel, K. Gartsman, I. Lubomirsky, J. Fleig, and J. Maier, "On the origin of the lattice constant anomaly in nanocrystalline ceria," *Phys. Chem. Chem. Phys.* **8**, 1111 (2006).
44. J. L. M. Rupp, A. Infortuna, and L. J. Gauckler, "Microstrain and self-limited grain growth in nanocrystalline ceria ceramics," *Acta Mater.* **54**, 1721 (2006).
45. S. Kim, R. Merkle, and J. Maier, "Oxygen nonstoichiometry of nanosized ceria powder," *Surf. Sci.* **549**, 196 (2004).
46. P. Patsalas, S. Logothetidis, L. Sygellou, and S. Kennou, "Structure-dependent electronic properties of nanocrystalline cerium oxide films," *Phys. Rev. B* **68**, 035104 (2003).
47. M. T. Ta, D. Briand, Y. Guhel, J. Bernard, J. C. Pesant, and B. Boudart, "Growth and structural characterization of cerium oxide thin films realized on Si(111) substrates by on-axis rf magnetron sputtering," *Thin Solid Films* **517**, 450 (2008).
48. T. Suzuki, I. Kosacki, and H. U. Anderson, "Microstructure-electrical conductivity relationships in nanocrystalline ceria thin films," *Solid State Ionics* **151**, 111 (2002).

Index

absorption, 7, 16, 54, 65, 67, 69, 70, 73, 85, 86, 91, 92, 294-298, 301, 331, 332, 335, 338-343, 346, 386, 395
  free-carrier, 332, 337, 339, 341, 343, 346
adhesion, 272, 375
adsorption, 363
AlGaAs, 172, 181, 338, 346, 388, 389
algorithm, 19, 334, 381, 382
aluminum, 66, 188, 205, 273, 275, 278, 279
amorphous silicon, 366
amplifier, 33, 47, 96, 407
antenna, 80, 81, 266
anti-crossing, 90, 92
antimony, 171
arsenic, 6, 7, 12, 29, 39, 43, 44, 47, 50, 51, 58, 72, 76, 77, 80, 81, 91, 97, 102, 107, 109, 112, 125, 127, 131, 132, 136, 146, 148, 163, 165, 166, 167, 169, 171, 178, 185, 194, 197, 199, 202, 206, 212, 213, 233, 241, 244, 251, 255, 259, 266, 267, 268, 270, 274, 300, 301, 303, 308, 309, 310, 326, 337, 340, 341, 344, 349, 351, 356, 357, 359, 368, 371, 378, 383, 384, 389, 396, 401, 403, 404, 412, 414, 416, 417, 418
atomic force microscopy (AFM), 9
atomic layer deposition (ALD), 146, 204, 421

ballistic transport, 113, 114, 141, 192, 294, 368, 369
band structure, 135, 290, 352, 361, 388
bandgap engineering, 294
bandwidth, 5, 7, 15, 16, 23, 25, 29, 31, 33, 34, 37, 39, 43, 81, 136
bit error rate (BER), 44
Boltzmann constant, 296
Boltzmann transport equation, 162
brain, 265, 266, 267, 268, 269, 270, 271, 272
breakdown voltage, 202
buffer layer, 110, 239

capacitance, 46, 80, 121, 131, 132, 134, 136, 140, 191, 212, 304
carbon, 107, 127, 131, 134, 139, 151, 202, 242, 245, 293, 365, 366, 374, 375, 376, 379, 381, 411
carrier leakage, 49, 50, 52, 53, 62
CdZnTe, 321, 322, 323, 324, 325, 326, 327, 328, 329, 346, 361, 362
cell library, 238
cellular automata, 130, 138
chalcogenide semiconductors, 107, 219, 220, 222, 223, 225, 228
chemical vapor deposition (CVD), 183, 184, 186, 187, 189, 204, 205
circular polarization, 255
clock distribution, 233, 235
clock frequency, 232, 235, 238
clock synchronization, 232, 233, 234, 235, 236
clocking, 235
CMOS, 7, 12, 16, 18, 30, 47, 109, 112, 113, 117, 119, 123, 124, 127, 129, 130, 132, 133, 135, 138, 139, 141, 143, 144, 146, 148, 149-152, 156, 169, 171, 179, 184, 187, 189, 191, 192, 210, 218, 232, 234, 237, 281
  scaling, 31, 43, 81, 109, 113, 117-119, 122, 126, 129, 131, 132, 135, 139, 157, 159, 160, 169, 170, 171, 198, 201, 211-213, 224, 244, 281, 368, 369, 370, 376
complex materials, 19
Compton telescope, 333-345, 347

Future Trends in Microelectronics. Edited by Serge Luryi, Jimmy Xu, and Alex Zaslavsky

Copyright © 2010 John Wiley & Sons, Inc.

computer, 29, 33, 229, 230, 232, 270, 282
conduction band, 62, 76, 113, 119, 151, 171, 175, 178, 194, 282, 283, 284, 286, 290, 294, 352
conductivity, 160, 161, 162, 163, 169, 171, 176, 178, 179, 196, 222, 226, 227, 282, 296, 298, 299, 300, 302, 305, 412, 413, 414, 415, 418, 419, 421, 422
contact resistance, 112
copper, 29, 43, 51, 66, 89, 90, 278, 290, 413
correlation effects, 358, 359, 361
Coulomb blockade, 186, 202
crazy carpentry, 352
crosstalk, 29
current drive, 46, 112, 113, 115-117, 119, 152
current oscillations, 273

dense wavelength division multiplexing (DWDM), 33
depletion, 58, 59, 120, 387
detector
  gamma radiation, 323, 326, 328
  infared (IR), 52, 392, 393
device area, 55
diblock copolymer, 183, 184, 187
dielectric, 6, 7, 9, 10, 18, 19, 20, 21, 37, 39, 47, 57, 58, 80, 86, 91, 92, 112, 113, 117, 126, 132, 163, 165, 169, 172, 190, 202, 203, 204, 206, 209, 210, 212, 224, 244, 281, 311, 332
  high-κ, 116, 119, 125, 126, 154, 155, 160, 169, 172, 204, 209, 210, 244, 245, 281, 290
diffraction limit, 37, 38, 54, 62
diffusion coefficient, 256, 343
dislocations, 143
display, 90, 188
  flat panel, 365
distributed feedback, 77, 83, 84
DNA, 75
domain, 5, 81, 95, 117, 136, 163, 251, 276, 282, 373, 402, 403, 404
doping profile, 58, 180
drain current, 114, 116, 121, 148, 165, 167, 180, 202, 241, 242, 244, 368, 369, 370
drain-induced barrier lowering (DIBL), 113, 114, 115, 118, 152, 165
drift-diffusion, 162, 165, 166, 253
dynamic RAM. See memory

effective mass, 65, 110, 134, 171, 173, 175, 192, 194, 197, 282-284, 286, 287
efficiency, 8, 33, 40, 49-52, 62, 65, 68-72, 75, 79, 80, 81, 107, 112, 129, 130, 148, 234, 236, 237, 261, 294, 321-325, 331, 332, 333, 335, 338-343, 366, 371, 373, 375, 376, 384, 386, 411
  injection, 69, 71, 305
electric field, 80, 99, 120, 136, 138, 144, 151, 160, 170, 179, 192, 194, 196, 201, 202, 204, 206, 207, 208, 220, 222, 225, 226, 228, 253, 254, 256, 257, 259, 293, 294, 297-304, 311-316, 371, 388, 389, 390, 391
electroluminescence, 45, 46, 59, 62
electron spin, 255, 258, 262
electron temperature, 388
electrooptic detection, 93
energy branching, 344, 347, 348, 349, 351, 352, 353, 354, 355, 356, 357, 358, 361, 363
energy conversion, 365
energy resolution, 321, 322, 323, 324, 325, 326, 328, 332, 344, 345, 347, 348, 356
epitaxy, 87, 143, 186, 315
  chemical beam epitaxy (CBE), 307, 313
  chemical beam epitaxy (MBE), 307

- molecular beam epitaxy (MBE), 66, 87, 89, 154, 175, 177, 180, 181, 239
- evolution, 20, 107, 127, 132, 144, 319, 353, 354, 373, 402, 413, 414, 417
- excitons, 139, 311, 312, 313, 315, 371, 372, 373

- fabrication technologies, 189
- Fano factor, 347, 348, 349, 350, 351, 352, 353, 355, 357, 358, 359, 360, 361, 362, 363
- feature size, 6, 9, 10, 11, 14, 18, 37, 109, 211, 215
- Fermi level pinning, 181
- ferroelectricity, 127
- fin-FET. See transistor
- flexible electronics, 366
- flip-flop, 233
- fluctuations, 84, 92, 117, 125, 159, 169, 176, 181, 192, 196, 234, 236, 241, 242, 243, 245, 282, 350, 351, 357, 358, 361, 362, 386
- fluorescence, 348, 350
- focused ion beam (FIB), 97, 98, 103, 104
- foundry, 234
- Fourier transform infrared (FTIR) spectrometer, 90
- four-wave mixing, 97, 102, 103
- frequency domain, 5
- fuel cells, 411, 413, 419-421
  - solid oxide, 419-421

- GaAs, 63, 64, 76, 86, 89, 92, 104, 115, 116, 141, 172, 177, 181, 192, 198, 199, 200, 255, 262, 275, 315, 332, 338, 346, 363, 385, 388, 389, 392
- gain spectrum, 67, 71
- gamma radiation, 332-334, 338, 343
- gamma-ray imaging, 323, 326, 328
- GaN, 175, 180, 181, 241, 242, 245
- gate capacitance, 114, 115, 125, 128, 136
- gate delay, 29, 159, 234, 235, 236
- gate insulator, 117, 142, 144, 163
- gate oxide, 110, 146, 159, 188, 201
- gate stack, 116, 124, 125, 146, 147, 149, 155, 160, 172, 281
- germanium, 109, 111, 115, 116, 119, 134, 141-157, 169, 185, 186, 190, 191, 195, 196, 197, 200, 228, 321, 324, 347-350, 362, 363
- germanium-on-insulator (GeOI), 141-156, 191
- gold, 262, 321
- grain boundaries, 417, 422
- graphene, 134, 135-139, 243, 249, 290, 293, 294, 303, 305, 306

- hafnium oxide (HfO_2), 116, 124, 142, 146, 147, 155, 204, 206, 208, 210
- Hall effect, 251, 252, 254, 255, 256, 259, 261, 262, 275, 277
  - spin (SHE), 249, 251, 255, 256, 260, 261, 262, 263
- Hall effect measurement, 239
- heterointerface, 333, 373
- heterojunction, 135, 371, 375, 376
- heterostructure, 67, 68, 70, 71, 72, 86, 119, 120, 126, 139, 156, 193, 239, 371
  - AlGaAs/GaAs, 388
  - InAlAs/InGaAs, 239, 240, 245
  - InP/InGaAs, 315
  - Si/SiGe, 184, 191, 192, 193, 196, 197, 198, 282, 290
  - SiGe/Si, 192, 199, 200, 290
- hexagonal, 181, 183
- HgI_2, 321, 322, 325, 329
- hybrid technology multithreaded architecture (HTMT), 230, 237
- hydrogen, 58, 110, 142, 196, 411
- hysteresis, 202, 275-277, 279, 395, 396-399, 402, 404-409

- III-V compounds, 171

image processing, 5, 19
impact ionization, 61, 119, 121, 347-354, 361
implants, 155, 156
InAs, 63, 86, 96, 177, 178, 179, 180, 181, 192, 308, 309, 310, 311, 312, 313, 314, 315, 316, 388, 389, 392
infrared (IR) detector, 52, 392, 393
InGaAsP, 104, 333, 337, 345
InN, 175, 176, 177, 178, 179, 180, 181
InP, 52, 58, 59, 61, 239, 307-316, 332, 333, 335-340, 343, 345, 346, 392
input/output (I/O), 29, 41, 43
integrated circuit (IC), 7, 10, 16, 24, 46, 109, 123, 156, 189, 281, 323, 370, 374
integration, 5, 13, 14, 15, 16, 18, 36, 40, 42, 46, 47, 97, 109, 111, 113, 117, 118, 123, 125, 126, 129, 149, 150, 152, 153, 155, 156, 157, 171, 174, 179, 180, 183, 212, 216, 224, 269, 290, 333, 384
interconnect, 7, 16, 25, 29, 31, 48, 134, 150, 153, 290
  optical, 3, 13, 19, 23, 25, 29, 30, 31, 37, 38, 39, 41, 42, 43
Internet, 23, 34, 35, 36
Internet protocol (IP), 34, 35, 42
intersubband transition, 89, 90, 91, 96
inversion layer, 151, 283, 284, 287
ion implantation, 110, 124
ITRS roadmap, 42, 116, 117, 125

Josephson junctions, 81, 137, 231, 238
Joule heating, 161, 169, 219, 225, 398, 407

kinetic equation, 353

laser
  diode, 32, 35, 64
  mid-infrared, 65
  population inversion, 294, 302, 306
  quantum cascade (QCL), 62, 63, 75, 77, 80, 81, 83, 84
  quantum well, 104
  quantum-box, 63
  semiconductor, 14, 62, 63, 64, 73, 83
  terahertz (THz), 84
  tunable, 15, 100
  vertical cavity surface emitting (VCSEL), 44, 45, 46, 47, 48
lattice constant, 193, 194, 416, 417, 422
lattice mismatch, 171
LC circuit, 81, 82, 83
leakage current, 115, 116, 117, 123, 141, 159, 212, 281
lifetime, 63, 77, 97, 130, 338, 346, 385, 386, 387, 388, 389, 391
  carrier, 63, 386, 387
  photoelectron, 388, 389, 390, 391
lightly doped drain (LDD), 156
linewidth, 15
lithography, 7, 15, 55, 183, 192, 201, 204, 211, 212, 213, 214, 215, 217
  electron-beam, 55, 57, 172, 275
  next generation, 213
  optical, 211, 213, 214
  photo, 147, 188, 413
  x-ray, 213

magnetoresistance, 200, 275, 276, 277, 306
magnetotransport, 171, 175, 177, 179, 192
membranes, 47, 313, 314, 411, 416, 418
memory
  biological, 265-272
  dynamic random access (DRAM), 109, 121, 123, 124, 211, 218, 232

  flash, 122, 202, 211-218, 220, 223, 227, 270
  NAND, 211
  NOR flash, 201, 211
  one-transistor (1T), 123, 126
  phase-change random access (PRAM), 222, 224
  static random-access (SRAM), 109, 123, 124, 126, 150, 156, 216, 290
memory cell, 201, 202, 203, 204, 205, 206, 209, 210, 215, 218, 219, 221, 222
mesa, 47, 146, 172
metal oxide semiconductor (MOS), 113, 151, 155, 159, 170
metal-insulator transition, 136, 395
microelectronics, 183, 213, 219, 270, 282, 365, 366
microprocessor, 29, 43, 230, 231, 233, 237
mid infrared (MIR), 63, 65, 75, 81, 83, 86, 92, 93
mobility, 110, 111, 113, 117, 124, 125, 141, 144-150, 154, 155, 159, 160, 162, 171-173, 179, 192, 193, 196-198, 200, 202, 209, 222, 240, 253, 261, 281, 290, 331, 369, 373, 374, 389
  degradation, 159, 162
  model, 162
modulation, 11, 14, 15, 16, 17, 24, 31, 32, 36, 38, 39, 40, 41, 42, 45, 46, 47, 92, 120, 136, 139, 193, 194, 195, 196, 245
modulator
  electrooptic, 46
modules, 29, 32, 35, 36, 40
monolithic integration, 150
Monte Carlo simulation, 113, 114, 196, 200, 359, 362
Moss-Burstein shift, 332, 335, 339, 342
multiferroics, 140
multiple quantum well (MQW). See quantum well
multithreading, 231, 237

nanoelectronics, 109, 118, 124, 170, 179, 183, 186, 189, 191, 192, 290, 305
nanofabrication, 11, 18, 97
nanomaterials, 183
nanotube
  carbon (CNT), 119, 131, 134, 139, 151, 202, 242, 245, 366, 370, 374, 375, 376
nanowire, 110, 119, 125, 126, 131, 139, 157, 172, 174-178, 180, 181, 184, 190, 201-203, 209, 210, 242, 245, 367
national security, 229, 237
negative refraction, 14
neural network, 5, 19
neurons, 266, 267, 271
nitridation, 144, 145
nitrides, 181
noise, 29, 123, 239, 241, 242, 243, 244, 245, 324, 327, 347, 361, 381, 386, 390, 398, 407
  1/*f*, 239, 243, 244
  low-frequency, 240, 242, 244
nonvolatile memory, 126, 186, 190, 201, 202, 209, 210
numerical aperture (NA), 195, 213, 341

optical communications, 23, 24, 25, 29, 31, 32, 35, 37, 38, 39, 40, 41, 309
optical fiber, 21, 24, 29, 31, 32, 33, 35, 104
optical network, 25, 26, 27, 28, 29, 31, 35, 37, 40, 41, 42
optical pumping, 294, 295, 296, 297, 298, 300, 301, 305

package, 30, 98, 216

pair excitation energy, 347, 351, 352, 354, 356, 360, 361
parasitic capacitance, 131, 212
parasitic effects, 389
passivation, 58, 59, 62, 142, 144, 145, 153, 154, 185, 186, 196
percolation, 366, 367, 368, 369, 370, 371, 374, 402, 405, 407, 408
  heterogeneous, 369
  transition, 402, 405, 407, 408
permittivity, 132, 297
personal computer (PC), 211
petaflops, 230, 237
phase coherence, 176
phase transition, 138, 219, 220, 222, 223, 319, 395, 396, 404, 408, 409, 413
phonons, 160, 161, 162, 196, 197, 294, 301, 348, 350, 351, 360, 385, 389
  bottleneck, 52, 63, 385
  losses, 351
phosphorus, 155
photodetectors, 331, 333, 388, 392, 393
photodiodes, 35, 332, 333, 336, 337, 344
photon, 3, 39, 40, 65, 86, 87, 90, 91, 92, 93, 94, 95, 97, 102, 103, 294, 295, 307, 309, 313, 315, 316, 331, 332, 333, 334, 335, 338, 339, 340, 341, 342, 343, 344, 386, 387
  collection efficiency, 341, 343, 344
  recycling, 332, 335, 338, 340, 341, 342, 343, 344
photonic crystal, 3, 6, 14, 20, 22, 24, 39, 40, 55, 78, 313, 314, 316
photonics, 23, 24, 35, 36, 37, 40, 41, 190
photovoltaics, 3, 375
pinch-off, 368
planarization, 47, 150
plasma, 59, 97, 144, 175, 184, 185, 189
plasmonics, 38, 39, 42
platinum, 413, 421
Poisson equation, 194
polaritons, 86, 88, 89, 90, 91, 93, 94, 95, 96
polysilicon, 365
population inversion, 294, 302, 306
potential barrier, 386, 387, 388, 389, 391
potential well, 194
power dissipation, 138, 159
processor design, 230, 232, 233, 234
propagation delay, 37, 114, 115, 234

quantum box, 49, 63
quantum cellular automata (QCA), 128, 130, 131, 133, 136
quantum computation, 129, 249, 281
quantum confinement, 131, 148
quantum dot (QD), 44, 48, 63, 95, 172, 180, 185, 186, 189, 192, 307-316, 385, 387-393
quantum electrodynamics, 86, 95, 104
quantum entanglement, 313, 316
quantum interference, 130, 179
quantum key, 313
quantum well (QW), 49, 65, 66, 67, 71, 72, 73, 76, 86, 87, 89, 92, 96, 104, 172, 192, 197, 198, 200, 290, 388, 391
  multiple (MQW), 69, 70, 71, 87, 96
quantum wire, 7, 14, 15, 16, 18, 179, 181, 200
qubit, 95

radiation, 37, 39, 75, 77, 81, 91, 294, 295, 296, 297, 298, 305, 306, 323, 326, 327, 328, 329, 331, 332, 333, 337, 344, 345, 350, 352, 361, 362, 386, 387, 390, 413
random parking, 351, 352, 357

rapid single flux quantum (RSFQ), 127, 138, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238
recombination, 65, 66, 67, 69, 70, 71, 73, 301, 306, 312, 332, 335, 337, 341, 342, 345, 371, 376
  radiative, 301, 343, 345
refractive index, 9, 10, 11, 39, 78, 81, 297, 298, 299, 301, 333, 404
relaxation time, 49, 52, 61, 256, 257, 296, 300, 301
reliability, 44, 49, 52, 53, 191, 210, 236, 375
resistance, 112, 115, 119, 137, 148, 153, 156, 163, 167, 178, 179, 219, 223, 224, 225, 243, 245, 276, 286, 367, 368, 398, 399, 407, 408
resolution, 5, 6, 8, 12, 18, 21, 79, 90, 91, 93, 98, 213, 311, 322, 323, 325, 328, 329, 332, 344, 345, 347, 379, 381, 383
resonant tunneling, 73
response time, 379, 381, 382
rf performance, 112
ring resonator, 199
rotating wave approximation, 87

saturation, 148, 162, 173, 174, 176, 414
  velocity, 162
scattering, 16, 52, 62, 77, 84, 93, 160, 161, 162, 163, 169, 170, 172, 176, 179, 196, 197, 198, 243, 252, 257, 258, 259, 293, 306, 334, 347, 352, 369, 389, 392
Schottky, 109, 112, 171, 179, 204, 311, 312
  barrier, 112, 171, 179
  source-drain, 109, 112
Schrödinger equation, 175, 194, 196, 284
scintillators, 331-333, 335-341, 344-346, 362
self-assembled quantum dot (SAQD), 63, 315, 316, 392
self-organization, 180, 192
semiconductors
  III-V, 7, 16, 30, 47, 65, 73, 109, 119, 127, 134, 149, 171, 180, 348
  work function, 117
sensors, 14, 57, 62, 96, 183, 365, 379, 380, 381, 383, 384, 385
  network, 380, 382, 383, 384
series resistance, 112
sheet density, 51
short channel effects, 113, 115, 116, 160, 165
shot-glass model, 349, 351, 360
SiGe, 110, 119, 126, 143, 144, 154, 160, 169, 185, 189, 191, 192, 193, 195, 196, 197, 198, 199, 200, 282, 286, 290
silicon, 7, 11, 14, 16, 17, 19, 22, 29, 37, 39, 42, 43, 47, 97, 104, 107, 109, 112, 113, 119, 123, 127, 129, 135, 141, 142, 145, 146, 149, 154-157, 159-163, 165, 169, 170, 172, 180, 185-189, 201, 202, 204, 210, 216, 227, 232, 245, 249, 281-291, 362, 366, 367, 371, 374-376
  photonics, 7, 19, 29, 39, 42
silicon carbide (SiC), 294, 297, 300, 306
silicon dioxide (SiO_2), 11, 14, 17, 37, 57, 97, 98, 103, 104, 143, 146, 162, 163, 165, 172, 175, 184, 186, 187, 188, 193, 202, 203, 204, 205, 206, 208, 210, 275, 307, 308, 311, 396, 421
silicon nitride (Si_3N_4), 57, 58, 59, 60, 61, 97, 98, 99, 103, 104, 203, 412, 414, 415, 416, 417, 418
silicon-on-insulator (SOI), 7-18, 30, 39, 109-115, 117, 121, 123-126, 129, 141, 143, 144, 148-156, 159, 160, 163-170, 173, 180, 209, 291
  fully depleted, 148
  ultra-thin, 117, 198

single photon, 39, 40, 42, 85, 95, 309, 313
  source, 309, 313
single-electron transistor (SET). See transistor
software, 23, 33
solar cells, 3, 365, 366, 371, 372, 373, 375, 376
space charge, 194
spin, 127, 130, 133, 134, 177, 179, 180, 181, 192, 249, 251-263, 274, 278-280, 282, 283, 290, 311, 312
  accumulation, 251, 252, 256, 257, 262
  current, 251, 252, 253, 256, 257, 258, 259, 260, 261, 262, 263
  injection, 255, 259
  relaxation, 181, 192, 252, 254, 257, 258, 259
  transport, 134, 259, 261, 290
spin-orbit interaction, 130, 177, 251, 253, 254, 257, 258, 278, 282
spintronics, 128, 133, 138, 249, 261
spontaneous emission, 20, 81, 84
stored charge, 212
strain, 51, 66-68, 73, 110, 111, 117, 124, 150, 172, 173, 180, 192-194, 198, 200, 281, 283-290, 307, 402
  relaxation, 110, 124, 173, 198
strained layers, 196
subthreshold slope, 110, 118, 119, 126, 132, 134, 136, 156, 174
superconductors, 95, 127, 137, 229, 230, 232-238, 273-275, 277-280, 395, 398
superlattice, 63, 184, 185, 190, 275, 276, 277
switching, 33-36, 39, 86, 91, 94, 95, 120, 126, 127, 129-134, 136, 159, 180, 191, 219, 220, 225, 227, 228, 282, 404, 418

telecommunications, 3, 6-8, 309, 313
temperature
  growth, 187
  high, 46, 143, 150, 169, 220, 367, 371
  low, 148, 150, 176, 179, 186, 187, 188, 195, 196, 197, 202, 210, 262, 278, 313, 367, 417
  room, 49, 51, 57, 61, 62, 65-70, 72, 73, 83, 86, 90, 93, 119, 139, 142, 180, 194, 196-198, 202, 208, 232, 233, 261, 263, 278, 294, 301, 305, 324, 332, 335, 338, 341, 385-389, 391-393, 396, 398, 408, 412
temperature coefficient of resistance (TCR), 398-408
terahertz (THz), 25, 38, 75, 80-86, 93-95, 294-306, 392
thermal budget, 112
thermal conductivity, 160, 161, 162, 163, 169, 170
thermal resistance, 163, 166, 167, 168
thermal stability, 199
thin film transistor (TFT). See transistor
three-dimensional (3D) integration, 117, 150, 189, 213, 216
threshold, 52, 65, 66-73, 114, 116, 117, 122, 125, 142, 144, 146, 149, 159, 165, 180, 202, 208, 220, 224, 301, 331, 347-351, 359, 361, 369, 370
  voltage, 114, 116, 117, 120, 122, 125, 142, 144, 146-149, 159, 165, 180, 202, 208, 212, 215, 220, 224, 225, 227
time domain, 94, 95, 239
timing, 235
transconductance, 119, 165, 198
transfer characteristics, 110, 114, 120
transistor
  carbon nanotube (CNT), 375
  dual gate, 113-116, 119, 120, 122, 123, 160, 164-168

- field-effect (FET), 119-121, 126, 128, 131, 134-140, 151, 152, 156, 157, 172, 174, 180, 188, 198, 202, 245, 306
- fin-FET, 117, 118, 120, 123, 164-168
- four-gate, 114
- gate-all-around, 114, 116, 117, 125, 203, 210, 282
- metal gate, 109, 117, 125, 141, 155, 169, 170, 187, 281, 290
- metal oxide semiconductor (MOSFET), 111, 112, 113, 114, 119, 120, 125, 141, 154, 159, 169, 170, 186, 203, 210, 281, 283, 288
- modulation-doped (MODFET), 239, 240, 388, 393
- molecular, 139
- molecular (MT), 139
- multi-channel, 100, 109, 116, 125, 157, 323
- multi-gate (MuGFET), 109, 113, 114, 115, 117, 121, 123, 124, 125, 209, 281
- single-electron (SET), 180
- spin, 130
- thin film (TFT), 366, 367, 369, 370
- tunneling (TFET), 119, 120, 126, 139, 142, 151, 152, 156, 157, 174

transit time, 304, 390
transit-time device, 294, 302
transmission line, 236
trench, 121
tungsten, 156
tunnel (Esaki) diode, 134
tunneling, 115, 116, 119, 120, 125-127, 131, 134, 135, 137, 139, 142, 148, 151, 156, 157, 160, 201-207, 212, 224, 244, 294, 302-306, 387, 392
- correlated, 136, 137
- Fowler-Nordheim, 202
- interband, 119, 121, 127, 148
- junction, 134

two-dimensional electron gas (2DEG), 87, 175, 176, 178, 179, 192, 199, 200, 263, 284, 291

ultra large-scale integration (ULSI), 153, 155

vacuum radiation, 96
valence band, 67, 68, 69, 92, 119, 134, 142, 151, 192, 194, 195, 196, 294, 335, 348, 350, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361
- offset, 68

vanadium oxide (VO_2), 138, 319, 395, 396, 397, 398, 399, 401, 402, 404, 407, 408, 409
vapor-phase epitaxy (VPE), 63
variability, 116, 117, 125, 159, 366, 367
very large-scale integration (VLSI), 125, 141, 151, 156, 210, 218, 232, 238, 290
voxels, 333

wallplug efficiency, 52, 54, 56, 57, 62
wavefunction, 51, 284
waveguide
- metal, 78, 80
- polymer, 37
- silicon, 14
- slab, 6, 11

wavelength division multiplexing (WDM), 28, 31, 32, 33, 35, 40-42
wave-pipelining, 235
wide bandgap semiconductor, 321, 323, 325
wireless, 383

x-ray diffraction (XRD), 91, 413, 416, 417, 418

This page intentionally left blank